Protein Science (1997), 6:1661–1681. Cambridge University Press. Printed in the USA.

Bayesian statistical analysis of protein side-chain rotamer preferences

ROLAND L. DUNBRACK, JR.[1] AND FRED E. COHEN
Department of Cellular and Molecular Pharmacology, University of California, San Francisco,
San Francisco, California 94143-0450

(RECEIVED December 26, 1996; ACCEPTED April 28, 1997)

Abstract

We present a Bayesian statistical analysis of the conformations of side chains in proteins from the Protein Data Bank. This is an extension of the backbone-dependent rotamer library, and includes rotamer populations and average χ angles for a full range of ϕ,ψ values. The Bayesian analysis used here provides a rigorous statistical method for taking account of varying amounts of data. Bayesian statistics requires the assumption of a *prior distribution* for parameters over their range of possible values. This prior distribution can be derived from previous data or from pooling some of the present data. The prior distribution is combined with the data to form the *posterior distribution*, which is a compromise between the prior distribution and the data. For the χ_2, χ_3, and χ_4 rotamer prior distributions, we assume that the probability of each rotamer type is dependent only on the previous χ rotamer in the chain. For the backbone-dependence of the χ_1 rotamers, we derive prior distributions from the product of the ϕ-dependent and ψ-dependent probabilities. Molecular mechanics calculations with the CHARMM22 potential show a strong similarity with the experimental distributions, indicating that proteins attain their lowest energy rotamers with respect to local backbone–side-chain interactions. The new library is suitable for use in homology modeling, protein folding simulations, and the refinement of X-ray and NMR structures.

Keywords: Bayesian statistics; molecular mechanics; protein structure; rotamers; side chains

In the last 10 years, the Protein Data Bank (PDB) has more than tripled in size. As the PDB has grown, it has become possible to analyze protein structure in greater detail and with greater statistical certainty. Although some studies have been qualitative in nature, the size of the database now enables us to put some structural variables on a more sound statistical footing. One aspect of protein structure that has been analyzed over a period of many years is the distribution of protein side-chain conformations. Side-chain rotamer libraries (Chandrasekaran & Ramachandran, 1970; Cody et al., 1973; James & Sielecki, 1983; Ponder & Richards, 1987) consist of a list of discrete side-chain conformations and their associated probabilities determined from their frequency of occurrence in the PDB. In most cases, these conformations correspond to "rotamers" or local minima on potential energy maps (Bhat et al., 1979; Gelin & Karplus, 1979; Benedetti et al., 1983) with frequencies predictable from conformational analysis of organic molecules (Janin et al., 1978; Dunbrack & Karplus, 1994). The discreteness of rotamers is enforced by barriers of 4–10 kcal/mol due to the overlap of bond molecular orbitals in eclipsed conformations (Karplus & Parr, 1963).

As more high-resolution structures have become available in recent years, it has become possible to determine rotamer preferences as a function of backbone conformation. Earlier efforts found weak correlations of rotamer distributions in different secondary structures (Janin et al., 1978; McGregor et al., 1987; Sutcliffe et al., 1987; Schrauber et al., 1993). With an extended database of 132 protein chains with resolution better than or equal to 2.0 Å, we compiled a backbone-dependent rotamer library (Dunbrack & Karplus, 1993) that gives the side-chain χ_1 rotamer distribution for each amino acid type and each occupied 10° by 10° region of the ϕ,ψ conformation space of the backbone. χ_1 rotamer preferences show detectable patterns as a function of ϕ and ψ for all side chains that can be explained by simple steric conformational analysis (Dunbrack & Karplus, 1994). The backbone-dependent rotamer library has been shown to be useful as a tool for predicting side-chain conformations from backbone coordinates for homology modeling (Dunbrack & Karplus, 1993; Bower et al., 1997) and for NMR and X-ray structure refinement (Kuszewski et al., 1996). The results obtained are a significant improvement over backbone-independent rotamer libraries.

Although the larger database presently available affords a more complete view of the variation of rotamer preferences as a function of ϕ and ψ, much of the Ramachandran map is sparsely populated because of backbone–backbone steric exclusions. Even in well-populated regions of the Ramachandran map, some rotamers are quite rare and their frequencies are therefore statistically unreli-

Reprint requests to: Roland L. Dunbrack, Jr., Department of Cellular and Molecular Pharmacology, University of California, San Francisco, San Francisco, California 94143-0450; e-mail: dunbrack@cmpharm.ucsf.edu.

[1] Address after October 1, 1997: Institute for Cancer Research, Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA 19111.

able. For the purposes of homology modeling, protein folding simulations, experimental structure refinement, and comparison with energy calculations, we need more complete estimates of rotamer distributions than are available in the database. In this paper, we use Bayesian statistical analysis to account for the varying amount of information in the database for χ_1 backbone-dependent rotamer distributions. In addition, certain combinations of the χ_1, χ_2, χ_3, and χ_4 rotamers are rare and the Bayesian analysis provides a better estimate of their probability of occurrence than do the data alone. Such "sparse data corrections" have been used in the calculations of potentials of mean force from PDB data by Sippl (1990) and by Šali and Blundell (1993), although not labeled Bayesian statistics as such.

Briefly, Bayesian statistical analysis provides a framework for combining "prior" information about measurable quantities with (usually limited) experimental data to determine a better estimate of a parameter of interest than the data alone provide. A simple example is in the flipping of a coin that may be slightly biased. As prior information, perhaps after a visual inspection, we might postulate that the probability of heads (θ_{heads}) is fairly close to 1/2. We guess a probability distribution for θ_{heads} that extends between 0 and 1, but is peaked around $\theta_{heads} = 0.5$. If we flip the coin 10 times and get 7 heads, we would not be justified in saying that θ_{heads} is 0.7. We need to consider the likelihood of getting 7 heads from 10 tosses, given all possible values for the underlying parameter θ_{heads}. The likelihood function is denoted $p(y|\theta)$, because it is the probability of the data y given the parameter θ. In Figure 1, we show a prior distribution and likelihood function for the coin flipping example. It is clear that the likelihood of getting 7 heads from 10 tosses is significant even when $\theta_{heads} = 0.5$.

The goal of Bayesian analysis is a full view of the distribution of θ_{heads} given the data, in the coin flipping example, $p(\theta_{heads}|y_{heads},n)$. The central equation of Bayesian statistics, described in Materials and methods, combines the prior distribution and the likelihood function to reach the posterior distribution:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \qquad (1)$$

As shown in Figure 1, in the coin flipping example, the posterior distribution for θ_{heads} peaks just above 0.5 because of the observed data of 7 heads from 10 tosses. The extent of the shift from the data value (0.7) is incorporated into the analysis by the form of the prior distribution. In any case, as the number of observations, n, increases, the resulting distribution, $p(\theta_{heads}|y_{heads},n)$ becomes more concentrated at the observed ratio of heads to tosses.

It is central to Bayesian analysis that the posterior distribution is more than a point estimate for a parameter such as θ_{heads}; instead, it is a probability distribution over the full range of allowed values of the parameter. This aspect can be exploited by simulation, where values for the parameter can be drawn randomly from the posterior distribution and used to explore the distribution of any function of the parameter. This method is used here to show how well the calculated posterior distributions for rotamer probabilities correspond to the raw data in the PDB.

With a full analysis of the rotamer preferences of protein side chains from Bayesian statistics, we compare the experimental data with molecular mechanics calculations using the CHARMM potential. Previous calculations (Ponnuswamy & Sasisekharan, 1971; Sasisekharan & Ponnuswamy, 1971; Pullman & Pullman, 1974; Janin et al., 1978; Marcus et al., 1996) emphasized the limitations that individual side chains place on the backbone conformation. In contrast, we show the potential energy surfaces in a way that captures the limitations the backbone places on the side-chain conformation, and we compare the results with conformational analysis based on data on hydrocarbons such as butane and pentane, as described previously (Dunbrack & Karplus, 1994), and with the experimental distributions derived from the Bayesian analysis.

A number of protein-folding models suggests that backbone conformations are achieved earlier in folding than final side-chain conformations (Shakhnovich & Finkelstein, 1989; Bromberg & Dill, 1994; Dill et al., 1995). If this is the case, then the restrictions that the backbone places on side-chain conformations are of great importance in understanding the thermodynamics and kinetics of protein folding. In a subsequent paper, we will examine the correspondence between energy calculations and the experimental data in a more quantitative fashion with a statistical mechanical model that examines the influence of both local backbone conformation and tertiary packing interactions on protein side chains.

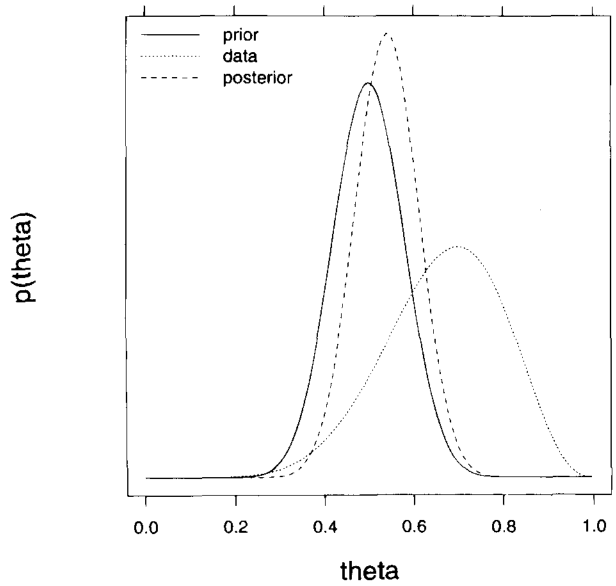

Fig. 1. Bayesian analysis of a set of coin flips. The prior density was calculated assuming 20 heads from 40 tosses for a perfect coin ($\theta_{heads} = 0.5$). The likelihood or data density was calculated assuming 7 heads from 10 tosses. The resulting posterior density is also plotted.

Results

Bayesian analysis of r_1-conditional backbone-independent rotamer populations

To make a clear distinction between the χ angles and their corresponding rotamers, we denote the χ_1 rotamer as r_1, the χ_2 rotamer as r_2, etc. The rotamer definitions for all side chains and χ angles are given in Table 1.

Given the strong dependence of the χ_1 rotamer probabilities on the backbone dihedrals ϕ and ψ, we present the backbone-independent rotamer library as the probabilities of χ_2, χ_3, χ_4 rotamers conditional on the χ_1 rotamer. The conditional backbone-independent rotamer library therefore consists of probability

Table 1. *Limits for rotamer library χ angles*

r_1 rotamers of all residue types except Pro r_2 rotamers of Arg, Gln, Glu, Ile, Leu, Lys, Met r_3 rotamers of Arg, Lys, Met r_4 rotamers of Arg, Lys		
r_1, r_2, r_3, r_4	Conformation	χ range
1	g^+	0° → 120°
2	t	120° → 240°
3	g^-	−120° → 0°
r_2 rotamers of Asn, Asp r_3 rotamers of Gln, Glu		
r_2, r_3	Conformation	χ range
1	g^+	30° → 90°
2	t	−30° → 30°
3	g^-	−90° → −30°
r_2 rotamers of Phe, Tyr, His		
r_2	Conformation	χ range
1	g	30° → 150°
2	t	−30° → 30°
r_2 rotamers of Trp		
r_2	Conformation	χ range
1	g^+	−180° → −60°
2	t	−60° → 60°
3	g^-	60° → 180°
r_1 rotamers of Pro		
r_1	Conformation	χ range
1	$g^{+(C_\gamma-\text{endo})}$	0° → 90°
2	$g^{-(C_\gamma-\text{exo})}$	−90° → 0°

distributions of $p(r_2,r_3,r_4|r_1)$. This is in contrast to traditional backbone-independent rotamer libraries which consist of values for $p(r_1,r_2,r_3,r_4)$ (e.g., Ponder & Richards, 1987).

For methionine, for instance, we would like to estimate values for the conditional probabilities of $r_2 = j$, $r_3 = k$ given that $r_1 = i$ or $p(r_2 = j, r_3 = k|r_1 = i)$ that would be measured from an infinitely large data set. We denote the infinite data set parameters for these quantities as $\theta_{jk|i}$. We derive prior distributions from a supposition that

$$p(r_2 = j; r_3 = k|r_1 = i) \propto p(r_3 = k|r_2 = j)p(r_2 = j|r_1 = i). \quad (2)$$

That is, rotamer probabilities along the chain of the side chain are dependent only on the previous dihedral. We use the probabilities in Equation 2 as the modes (maximum values) of prior distributions for each of the nine $\theta_{jk|i}$ for each r_1 rotamer i. The factors on the right-hand side of Equation 2 come directly from the raw data population probabilities for each side-chain type. Equation 2 is quite a good approximation with a correlation coefficient of 0.998 for the raw data probabilities versus the prior distribution estimate for Arg, Met, Gln, Glu, and Lys. The form of both the prior and posterior distributions is a Dirichlet function, a generalization of the multinomial distribution. In forming the posterior distribution from the prior distribution and the data, we need to choose how heavily to weight the prior distribution in relation to the data. An equivalent statement is that we need to define the variance of the prior distribution compared to the variance of the data. To accomplish this, we scale the prior distribution to correspond to some proportion of the data sample size with a scale parameter K. A value of 1.0 means that the prior "sample size" is equal to the data sample size, and the two parts then make equal contributions to the posterior distribution. We find a value of $K = 0.5$ to be reasonable, as described in the Materials and methods.

In Table 2, the backbone-independent prior distribution parameters derived from the raw data are listed for several side-chain types. These include values for $p(r_b|r_a)$, where $\{a,b\} = \{1,2\}$, $\{2,3\}$, or $\{3,4\}$, as well as the χ angle averages and their standard deviations, σ. The values for these parameters for the remaining side chains are available on our website (Dunbrack, 1997). The parameters from the posterior distribution for Met are listed in Table 3. Met has 27 $r_2,r_3|r_1$ rotamers, and 10 of these have been seen less than 10 times each in our set of 518 protein chains from the PDB. The posterior distribution provides a better estimate for the frequency of these rare rotamers than the data alone, because it contains very good estimates in the prior distribution. This is also the case for Arg and Lys, with 81 backbone-independent rotamers each, many of which have never been seen in protein structures because of their significant steric strain. It is the case that for all side chains except Asn and Asp, there is little or no dependence of $\theta_{j|i}$ or $\bar{\chi}_2(j|i)$ on ϕ and ψ. The backbone-independent library therefore consists of values for the parameters $\theta_{jkl|i}$ and $\{\bar{\chi}_2, \bar{\chi}_3, \bar{\chi}_4\}(j,k,l|i)$. The full backbone-independent library is also available on our website (Dunbrack, 1997).

Bayesian analysis of backbone-dependent rotamer populations

Determination of the backbone-dependent rotamer library is more difficult than the backbone-independent library because of the large number of parameters to be estimated. We would like to define the probabilities of the three r_1 rotamers (two for Pro) for all values of ϕ and ψ on a grid with 10° spacing. The infinite data parameters will be denoted $\theta_{i|ab}$ for the proportion of side chains with $r_1 = i$ in a region near $\{\phi = \phi_a, \psi = \psi_b\}$. Because of the constraint that $\Sigma_i \theta_{i|ab} = 1$, the total number of parameters is $2 \times 36 \times 36$ or 2,592 for each side-chain type. Because the data are very concentrated in particular ϕ,ψ regions (α-helix and β-sheet regions), we need some procedure for estimating the probabilities across the Ramachandran map. The full procedure is described in Materials and methods, but the results of each stage will be illustrated here.

The first step is to count the side chains of each rotamer type using weighting functions shown in Figure 2 (illustrated for $\phi_a = -60°$, $\psi_b = -60°$). The weighting function in Figure 2A, defined in Equation 7 in Materials and methods, has a Gaussian shaped peak at ϕ_a,ψ_b. It also has much smaller peaks at $\{\phi_a \pm 180°, \psi_b\}$ and $\{\phi_a, \psi_b \pm 180°\}$ and a very small peak at $\{\phi_a \pm 180°, \psi_b \pm 180°\}$. Because the backbone-conformation dependent steric interactions that affect rotamer preferences are periodic in ϕ and ψ every 180° (Dunbrack & Karplus, 1994) (described in detail below), the secondary peaks in the weighting function serve to supplement very sparse regions with data from regions of the Ramachandran map 180° away in ϕ and/or ψ. In most cases, heavily populated regions in the primary peak correspond to sparsely populated regions in the secondary peaks and vice versa. In Figure 2B, we show the function used in our previous work (Dunbrack & Karplus, 1993), denoted $W_{ab}^{box}(\phi,\psi)$, which gives weights of 1.0 to side chains with ϕ and ψ within 10° of the ϕ,ψ point and 0.0 to all others. In Figure 2C, the product of the functions in Figure 2A and B is shown. This function, denoted $W_{ab}^{non\text{-}per}(\phi,\psi)$, counts only side chains within 10° of ϕ,ψ, but with

Table 2. *Backbone-independent prior distribution parameters*[a]

Residue	$r_b \mid r_a$	r_a	r_b	$n(r_b \mid r_a)$	$p(r_b \mid r_a)$	$\bar{\chi}_a$	σ_a	$\bar{\chi}_b$	σ_b
Asn	$r_2 \mid r_1$	1	*1*	*229*	*27.86*	*59.1*	*(12.3)*	*59.3*	*(16.9)*
			2	**393**	**47.81**	**64.7**	**(8.3)**	**0.0**	**(16.1)**
			3	*200*	*24.33*	*66.0*	*(14.0)*	*−56.6*	*(17.0)*
		2	**1**	**639**	**41.87**	**-170.6**	**(14.3)**	**55.1**	**(16.3)**
			2	**611**	**40.04**	**-170.2**	**(13.3)**	**1.5**	**(17.7)**
			3	*276*	*18.09*	*−169.1*	*(14.9)*	*−57.1*	*(18.4)*
		3	*1*	*263*	*9.32*	*−75.9*	*(20.2)*	*68.1*	*(17.3)*
			2	**945**	**33.50**	**-72.6**	**(10.1)**	**-13.0**	**(12.8)**
			3	**1,613**	**57.18**	**-66.6**	**(12.7)**	**-56.6**	**(16.4)**
Leu	$r_2 \mid r_1$	1	*1*	*80*	*55.94*	*60.9*	*(17.4)*	*77.8*	*(18.4)*
			2	*59*	*41.26*	*67.1*	*(16.4)*	*164.3*	*(20.5)*
			3	*4*	*2.80*	*52.8*	*(12.1)*	*−42.3*	*(34.2)*
		2	**1**	**2,408**	**84.43**	**-176.5**	**(14.5)**	**63.6**	**(11.9)**
			2	*349*	*12.24*	*−156.3*	*(18.9)*	*−176.6*	*(29.4)*
			3	*95*	*3.33*	*−165.1*	*(16.3)*	*−75.1*	*(25.0)*
		3	*1*	*742*	*13.27*	*−93.3*	*(15.4)*	*42.7*	*(26.0)*
			2	**4,677**	**83.64**	**-65.5**	**(11.2)**	**175.8**	**(11.1)**
			3	*173*	*3.09*	*−82.7*	*(15.4)*	*−43.3*	*(24.9)*
Phe	$r_2 \mid r_1$	1	**1**	**541**	**98.90**	**62.5**	**(10.6)**	**91.0**	**(10.6)**
			2	*6*	*1.10*	*68.6*	*(11.7)*	*−2.6*	*(37.1)*
		2	**1**	**1,360**	**93.34**	**-178.7**	**(11.3)**	**77.1**	**(13.9)**
			2	*97*	*6.66*	*−172.5*	*(11.3)*	*28.4*	*(18.9)*
		3	**1**	**1,984**	**84.64**	**-66.4**	**(11.4)**	**97.6**	**(16.4)**
			2	*360*	*15.36*	*−70.3*	*(11.9)*	*−14.3*	*(23.1)*
Trp	$r_2 \mid r_1$	1	**1**	**156**	**58.43**	**60.8**	**(12.3)**	**-90.8**	**(10.6)**
			2	*15*	*5.62*	*69.4*	*(6.6)*	*−16.0*	*(42.8)*
			3	**96**	**35.96**	**58.9**	**(13.2)**	**88.9**	**(9.6)**
		2	**1**	**245**	**45.20**	**-177.8**	**(12.6)**	**-104.2**	**(14.9)**
			2	*107*	*19.74*	*−174.6*	*(12.7)*	*22.6*	*(28.2)*
			3	**190**	**35.06**	**-179.5**	**(10.8)**	**83.2**	**(10.5)**
		3	**1**	**94**	**11.42**	**-67.8**	**(14.3)**	**-90.8**	**(17.9)**
			2	*185*	*22.48*	*−69.2*	*(9.5)*	*−3.0*	*(26.9)*
			3	**544**	**66.10**	**-68.3**	**(10.6)**	**98.4**	**(16.5)**
Met	$r_2 \mid r_1$	1	*1*	*11*	*6.25*	*62.5*	*(18.5)*	*78.5*	*(11.9)*
			2	**157**	**89.20**	**61.0**	**(12.4)**	**-177.8**	**(15.2)**
			3	*8*	*4.55*	*67.4*	*(13.5)*	*−83.2*	*(12.3)*
		2	**1**	**185**	**30.43**	**-170.5**	**(14.8)**	**67.1**	**(15.6)**
			2	**380**	**62.50**	**-176.2**	**(15.1)**	**177.3**	**(14.8)**
			3	*43*	*7.07*	*−172.6*	*(17.3*	*−87.7*	*(12.4)*
		3	*1*	*22*	*1.71*	*−78.0*	*(27.6)*	*83.8*	*(20.9)*
			2	**754**	**58.72**	**-68.8**	**(11.1)**	**-178.8**	**(12.9)**
			3	**508**	**39.56**	**-65.7**	**(11.6)**	**-63.9**	**(13.4)**
	$r_3 \mid r_2$	1	**1**	**136**	**62.39**	**67.8**	**(15.4)**	**72.5**	**(16.9)**
			2	**55**	**25.23**	**71.6**	**(18.2)**	**-175.6**	**(30.6)**
			3	*27*	*12.39*	*72.9*	*(20.7)*	*−92.0*	*(27.3)*
		2	**1**	**542**	**41.98**	**179.1**	**(13.3)**	**71.9**	**(19.4)**
			2	**313**	**24.24**	**179.5**	**(14.7)**	**-176.0**	**(25.5)**
			3	**436**	**33.77**	**-177.9**	**(14.0)**	**-75.0**	**(19.5)**
		3	*1*	*74*	*13.24*	*−66.1*	*(13.6)*	*99.2*	*(12.5)*
			2	**99**	**17.71**	**-67.9**	**(16.9)**	**169.6**	**(28.3)**
			3	**386**	**69.05**	**-65.5**	**(14.5)**	**-70.4**	**(15.8)**

(continued)

Table 2. *Continued*

Residue	$r_b \mid r_a$	r_a	r_b	$n(r_b \mid r_a)$	$p(r_b \mid r_a)$	$\bar{\chi}_a$	σ_a	$\bar{\chi}_b$	σ_b
Arg	$r_2 \mid r_1$	1	*1*	*21*	*5.05*	*54.8*	*(25.3)*	*88.3*	*(14.3)*
			2	**382**	**91.83**	**63.1**	**(15.1)**	**-179.4**	**(18.9)**
			3	*13*	*3.12*	*65.3*	*(27.4)*	*-80.6*	*(22.5)*
		2	**1**	**245**	**15.69**	**-176.5**	**(17.2)**	**71.3**	**(19.2)**
			2	**1,251**	**80.09**	**-174.5**	**(15.4)**	**178.8**	**(20.0)**
			3	*66*	*4.23*	*-160.7*	*(22.8)*	*-86.9*	*(21.4)*
		3	*1*	*108*	*4.16*	*-85.4*	*(19.5)*	*77.9*	*(23.5)*
			2	**1,963**	**75.67**	**-67.9**	**(13.7)**	**-178.7**	**(18.3)**
			3	**523**	**20.16**	**-62.1**	**(14.7)**	**-72.5**	**(17.9)**
	$r_3 \mid r_2$	1	**1**	**121**	**32.35**	**77.6**	**(20.9)**	**69.8**	**(19.4)**
			2	**222**	**59.36**	**72.0**	**(20.5)**	**177.7**	**(24.6)**
			3	*31*	*8.29*	*76.0*	*(21.2)*	*-83.7*	*(21.5)*
		2	**1**	**915**	**25.44**	**176.8**	**(20.4)**	**67.1**	**(19.1)**
			2	**1,593**	**44.30**	**179.5**	**(17.5)**	**-179.2**	**(19.6)**
			3	**1,088**	**30.26**	**-175.5**	**(19.0)**	**-68.6**	**(19.1)**
		3	*1*	*68*	*11.30*	*-80.7*	*(22.2)*	*79.9*	*(22.0)*
			2	**316**	**52.49**	**-72.1**	**(16.9)**	**-179.4**	**(18.9)**
			3	**218**	**36.21**	**-75.4**	**(20.1)**	**-69.5**	**(19.1)**
	$r_4 \mid r_3$	1	**1**	**352**	**31.88**	**66.5**	**(17.9)**	**84.4**	**(17.4)**
			2	**631**	**57.16**	**68.4**	**(19.9)**	**-171.5**	**(29.0)**
			3	*121*	*10.96*	*72.0*	*(21.9)*	*-99.2*	*(22.5)*
		2	**1**	**467**	**21.91**	**176.7**	**(20.3)**	**86.8**	**(19.8)**
			2	**1,074**	**50.40**	**-179.1**	**(20.7)**	**177.5**	**(28.3)**
			3	**590**	**27.69**	**-177.3**	**(18.3)**	**-87.5**	**(17.4)**
		3	*1*	*179*	*13.39*	*-74.1*	*(22.2)*	*103.5*	*(13.2)*
			2	**688**	**51.46**	**-69.5**	**(19.3)**	**170.8**	**(29.4)**
			3	**470**	**35.15**	**-66.5**	**(17.6)**	**-86.4**	**(16.4)**

[a]Rotamer pairs with syn-pentane (1,3 or 3,1) interactions are in *italics*. All others are in **bold**. Rotamer designations are defined in Table 1.

a weight similar to that in Figure 2A. It will be used to count data for the likelihood function (see below). To check that the periodic and nonperiodic functions do not alter the probabilities significantly from our previous "box" function, we calculated the correlation coefficients for values of $\{\phi,\psi\}$ with more than 20 side chains counted with W^{box}. For W^{per}, the correlation coefficient with W^{box} was 0.987, and for $W^{non\text{-}per}$, the correlation coefficient with W^{box} was 0.999. Between W^{per} and $W^{non\text{-}per}$ the correlation coefficient was 0.985.

The next step is to use the weighted counts of W^{per} to derive a prior density distribution for use in a Bayesian analysis for the backbone-dependent rotamer library. We choose a prior density distribution centered on a product of ϕ-dependent densities and ψ-dependent densities, i.e.,

$$p(r_1 = i \mid \phi_a, \psi_b) \propto p(r_1 = i \mid \phi_a) p(r_1 = i \mid \psi_b). \qquad (3)$$

The factors on the right-hand side of Equation 3 can be calculated by taking the logs of each side of Equation 3 and solving the resulting linear equations by singular value decomposition. The approximation in Equation 3 is not quite as good as the backbone-independent case, but good enough to provide reasonable estimates of the probabilities throughout the Ramachandran map. For Arg, for instance, the correlation coefficient for $\{\phi,\psi\}$ values in occupied regions of the Ramachandran map (>20 side chains) between the prior distribution values calculated from the right-hand side of Equation 3 and the raw data values calculated with function $W^{non\text{-}per}$ is 0.952.

As with the backbone-independent prior distributions (Equation 2), values of $p(r_1 = i \mid \phi_a, \psi_b)$, denoted $\theta^{prior}_{i|ab}$, defined in Equation 3 indicate the central values (modes) of Dirichlet probability distributions or $p(\theta_{i|ab})$. In Figure 3, we show the prior distribution modes calculated with the weighting function W^{per} for several side-chain types plotted on top of bar charts of the data measured with the nonperiodic weighting function $W^{non\text{-}per}$ for values of ϕ_a and ψ_b with $y_a > 20$ or $y_b > 20$. The procedure used to derive the curves in Figure 3 attempts to decouple the ϕ and ψ dependence, and so, for instance, the ψ-dependence of the $r_1 = 3$ rotamer prior density is much flatter than the data density (the bars in Figure 3). In this case, the prior density does not respond to the strong variations of the probabilities of the other two rotamers with ψ.

In Figure 4, we show the prior distribution, the data density calculated with the weighting function W^{per} and with $W^{non\text{-}per}$, and the posterior distribution for the three r_1 rotamers of Arg. Comparing the data in $W^{non\text{-}per}$ with the other plots demonstrates

Table 3. *Methionine backbone-independent posterior distribution parameters*

r_1	r_2	r_3	$n(r_1)$	$n(r_1,r_2,r_3)$	$p(r_1,r_2,r_3)$	(σ)	$p(r_2,r_3\|r_1)$	(σ)	Ave.χ_1	$\sigma(\chi_1)$	Ave.χ_2	$\sigma(\chi_2)$	Ave.χ_3	$\sigma(\chi_3)$
1	*1*	*1*	*176*	*7*	*0.37*	*(0.11)*	*4.19*	*(1.21)*	*65.1*	*(22.4)*	*72.5*	*(6.9)*	*71.5*	*(17.7)*
		2	*176*	*3*	*0.17*	*(0.07)*	*1.97*	*(0.84)*	*56.8*	*(13.6)*	*88.2*	*(12.3)*	*147.4*	*(7.9)*
		3	*176*	*1*	*0.09*	*(0.05)*	*0.98*	*(0.60)*	*62.5*	*(18.5)*	*77.1*	*(10.3)*	*−92.0*	*(27.3)*
	2	**1**	**176**	**66**	**3.19**	**(0.31)**	**36.61**	**(2.91)**	**61.1**	**(10.1)**	**−177.2**	**(14.6)**	**75.3**	**(18.3)**
		2	**176**	**31**	**1.63**	**(0.23)**	**18.69**	**(2.36)**	**56.9**	**(17.5)**	**−177.7**	**(17.5)**	**179.3**	**(26.1)**
		3	**176**	**60**	**2.80**	**(0.29)**	**32.06**	**(2.82)**	**63.2**	**(11.4)**	**−178.5**	**(14.9)**	**−74.2**	**(20.5)**
	3	*1*	*176*	*0*	*0.05*	*(0.04)*	*0.56*	*(0.45)*	*67.4*	*(13.5)*	*−75.5*	*(9.1)*	*99.2*	*(12.5)*
		2	*176*	*3*	*0.15*	*(0.07)*	*1.72*	*(0.79)*	*62.0*	*(15.0)*	*−81.2*	*(18.4)*	*179.0*	*(36.4)*
		3	*176*	*5*	*0.28*	*(0.09)*	*3.21*	*(1.06)*	*71.2*	*(12.4)*	*−81.6*	*(8.8)*	*−62.7*	*(28.1)*
2	**1**	**1**	**608**	**119**	**5.68**	**(0.41)**	**19.29**	**(1.30)**	**−172.0**	**(12.8)**	**65.7**	**(13.8)**	**73.6**	**(15.4)**
		2	**608**	**45**	**2.22**	**(0.26)**	**7.53**	**(0.87)**	**−169.0**	**(14.8)**	**67.1**	**(15.6)**	**−171.3**	**(28.7)**
		3	*608*	*21*	*1.07*	*(0.18)*	*3.63*	*(0.62)*	*−165.6*	*(22.7)*	*74.7*	*(22.6)*	*−86.8*	*(29.0)*
	2	**1**	**608**	**163**	**7.79**	**(0.48)**	**26.47**	**(1.45)**	**−177.8**	**(14.5)**	**176.7**	**(14.3)**	**71.7**	**(20.7)**
		2	**608**	**89**	**4.35**	**(0.36)**	**14.77**	**(1.17)**	**−173.9**	**(15.7)**	**176.1**	**(16.5)**	**178.4**	**(26.4)**
		3	**608**	**128**	**6.17**	**(0.43)**	**20.97**	**(1.34)**	**−175.8**	**(15.4)**	**179.0**	**(14.4)**	**−75.0**	**(21.2)**
	3	*1*	*608*	*1*	*0.15*	*(0.07)*	*0.53*	*(0.24)*	*−172.6*	*(17.3)*	*−77.9*	*(9.1)*	*99.2*	*(12.5)*
		2	*608*	*7*	*0.38*	*(0.11)*	*1.28*	*(0.37)*	*−177.0*	*(28.4)*	*−91.0*	*(15.2)*	*−178.2*	*(30.2)*
		3	*608*	*35*	*1.62*	*(0.23)*	*5.52*	*(0.75)*	*−172.0*	*(14.4)*	*−86.6*	*(11.5)*	*−77.2*	*(11.4)*
3	*1*	*1*	*1,284*	*10*	*0.57*	*(0.13)*	*0.92*	*(0.22)*	*−91.6*	*(17.1)*	*86.4*	*(20.6)*	*61.6*	*(26.9)*
		2	*1,284*	*7*	*0.34*	*(0.10)*	*0.56*	*(0.17)*	*−57.7*	*(34.4)*	*89.0*	*(17.8)*	*174.7*	*(38.0)*
		3	*1,284*	*5*	*0.24*	*(0.09)*	*0.38*	*(0.14)*	*−83.7*	*(10.7)*	*62.5*	*(10.0)*	*−109.2*	*(7.4)*
	2	**1**	**1,284**	**313**	**15.09**	**(0.64)**	**24.41**	**(0.98)**	**−69.7**	**(11.0)**	**179.5**	**(12.2)**	**71.3**	**(19.0)**
		2	**1,284**	**193**	**9.12**	**(0.51)**	**14.75**	**(0.81)**	**−67.9**	**(10.8)**	**−179.4**	**(13.1)**	**−172.6**	**(24.5)**
		3	**1,284**	**248**	**12.03**	**(0.58)**	**19.45**	**(0.90)**	**−68.2**	**(11.4)**	**−176.2**	**(13.4)**	**−75.2**	**(18.4)**
	3	*1*	*1,284*	*73*	*3.44*	*(0.33)*	*5.56*	*(0.52)*	*−66.3*	*(10.5)*	*−66.0*	*(13.7)*	*99.2*	*(12.6)*
		2	**1,284**	**89**	**4.31**	**(0.36)**	**6.98**	**(0.58)**	**−65.8**	**(12.2)**	**−65.3**	**(14.9)**	**168.1**	**(27.8)**
		3	**1,284**	**346**	**16.70**	**(0.67)**	**27.00**	**(1.01)**	**−65.6**	**(11.7)**	**−63.1**	**(12.9)**	**−69.9**	**(15.8)**

the utility of the Bayesian method for combining an informative prior distribution with the $W^{non\text{-}per}$ data to produce the posterior distribution. The $W^{non\text{-}per}$ data cover a small percentage of the total ϕ,ψ plot with isolated regions with existent data and large regions with no data. The prior distribution is a smoothed and partially symmetrized reconstruction of the data. Adding back the nonperiodic data to the prior distribution to produce the posterior distribution ensures that, in populated regions, the posterior distribution represents the data quite closely. In less populated regions, however, the prior distribution is more important and more likely to be accurate than the nonperiodic data alone. In unpopulated regions, the raw data are of no use, and we must rely on the accuracy of the prior distribution defined in Equation 3.

In Figure 5, we show the results for three more side-chain types (Asp, Phe, and Val) to demonstrate the similarities and differences between various types of residues. Side-chain pairs with similar stereochemistries behave similarly, especially with similar chemistry at both the β and γ positions (e.g., Asp-Asn, Lys-Arg, Phe-Tyr). The differences between the β-unbranched and β-branched side-chain distributions are clearly due to Cγ/backbone interactions. Branching at γ and electrostatic interactions also affect the distributions in less drastic ways (e.g., Asp-Arg, Lys-Phe).

As noted earlier, the backbone-dependent library does not include a dependence of r_2 rotamers on the backbone dihedrals ϕ and ψ, except in the case of Asp and Asn. Asp and Asn χ_2 distributions are not described easily as rotamers, because the distributions are nearly continuous with only one mode. But there is a significant skew to the distributions due to backbone-conformation-independent syn-pentane effects with backbone N_i and C_i and backbone-conformation-dependent electrostatic effects with backbone atoms of residues $i - 1$, i, and $i + 1$. To represent the skewness of the distributions, we define rotamers for χ_2 for Asp and Asn in Table 1 centered at +60°, 0°, and −60°. With these definitions, we can plot the r_2 probabilities varying with ϕ and ψ.

There is a weak dependence of $p(r_2|r_1)$ on ϕ and ψ for Lys, Met, Glu, Gln, Arg, Ile, and Leu (data not shown). For example, the values of $p(r_2|r_1)$ for $r_2 = 2(t)$ range from 60% to 80% with ϕ and ψ in an identical manner in all of these side chains, with similar variations in the $r_2 = 1$ and $r_2 = 3$ rotamers. But for weakly populated values of ϕ and ψ, there are not enough data to determine these values, even with the periodic weighting function used to determine the backbone-dependence of the r_1 rotamers.

The situation for Asp and Asn is quite different. There are large shifts in $p(r_2|r_1)$ probabilities with changes in backbone dihedrals, in some cases shifts of density of 80% or more. The backbone-dependent rotamer library for Asp and Asn therefore contains values for $p(r_1,r_2|\phi,\psi)$, whereas for the other side chains, the backbone-dependent part of the library contains only the parameters for $p(r_1|\phi,\psi)$.

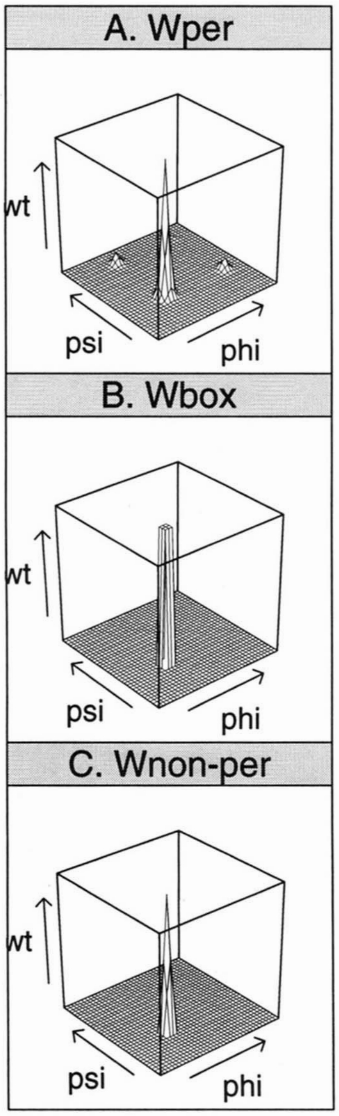

Fig. 2. Weighting functions for counting backbone-dependent rotamers. **A:** W_{ab}^{per} (ϕ,ψ); **B:** W_{ab}^{box} (ϕ,ψ); **C:** $W_{ab}^{non\text{-}per}$ (ϕ,ψ) for $\phi_a = -60°$, $\psi_b = -60°$. The functional forms are given in the text.

χ angles in side chains depend on a number of variables, but we would like to estimate their averages conditional upon backbone conformation. With an infinitely large database, we would simply calculate the distribution of χ angles, $\{\bar{\chi}_1, \bar{\chi}_2, \bar{\chi}_3, \bar{\chi}_4\}$, as a function of $\{\phi, \psi, r_1, r_2, r_3, r_4\}$. There are not enough data in the database for rare and even not-so-rare rotamers to determine this many parameters, and so we make use of informative prior distributions and calculate only a subset of the possible χ angle parameters. The prior distributions are based on the approximation that χ angle averages are likely to depend only on their own rotameric state and the rotameric states (or ϕ and ψ dihedrals in the case of r_1) one previous in the chain and one after in the chain. As an example, we show the one-dimensional prior distributions for χ_1 averages as a function of ψ in Figure 6 for all several side-chain types. The forms of these functions are explained easily with the conformational analysis reviewed in the next section. We combine the prior distributions for the χ_1 angles as shown in Figure 6 with the data to derive posterior distributions for the backbone-dependent averages of χ_1, The backbone-dependent library now consists of values for $\theta_{i|ab}$ and $\{\bar{\chi}_1\}(r_1 = i|\phi_a, \psi_b)$. These results can be combined with $\theta_{jkl|i}$ and $\{\bar{\chi}_2, \bar{\chi}_3, \bar{\chi}_4\}(r_2 = j, r_3 = k, r_4 = l|r_1 = i)$ to form the complete library.

Conformational analysis of backbone-independent interactions

As a very simple model for steric interactions in biopolymers, we have previously compared the distribution of side-chain rotamers with steric interactions that are well-documented in simple hydrocarbons such as butane and pentane (Dunbrack & Karplus, 1994). The barrier to rotation in ethane was estimated to be 3.0 kcal/mol as early as the 1930's (Kemp & Pitzer, 1937) with potential energy maxima in the "eclipsed" conformations (i.e., with a dihedral H-C-C-H = 0°) and minima with the hydrogens "staggered" with each H-C-C-H dihedral equal to +60°, 180°, or −60°. From spectroscopic data and the experimental thermodynamics of longer hydrocarbons, it was also clear that certain combinations of neighboring dihedrals, namely g^+, g^- and g^-, g^+, in longer hydrocarbons were not allowed because of steric hindrance between carbon atom i and $i + 4$ of the chain (Pitzer, 1940a, 1940b). Recent high-level ab initio calculations (Wiberg & Murcko, 1988) and experimental measurements (Durig & Compton, 1979; Compton et al., 1980) have shown that the single gauche interaction in butane is about 0.9 kcal/mol higher than the global minimum *trans* conformation (180°). The ab initio energy of two consecutive gauche interactions of like sign in pentane has an energy of 1.4 kcal/mol above the global minimum $\{t,t\}$ conformation, whereas the syn-pentane $\{g^+, g^-\}$ conformation had an energy of 3.3 kcal/mol above $\{t,t\}$ (Wiberg & Murcko, 1988).

Side-chain γ and δ heavy atoms (of the ith residue) can interact with backbone atoms N_i and C_i in ways that are dependent on the values of χ_1 and χ_2, These interactions have been recognized for some time as being responsible for the $\{r_1, r_2\}$ rotamer distribution of protein side chains (Chandrasekaran & Ramachandran, 1970). In Figure 7A, we show a Newman projection of $\{r_1, r_2\}$ rotamers for hydrocarbon side chains (Lys, Arg, Met, Glu, Gln, Leu, Val, Ile). In Table 4, the backbone-independent gauche and syn-pentane interactions for all side-chain types are listed. We can count these interactions to get a rough estimate of steric and dihedral strain in the various rotamer combinations. Each gauche interaction, g, costs about 0.9 kcal/mol, whereas each syn-pentane interaction costs $2g + p$, where $p = 1.5$ kcal/mol. We also include the CHARMM energies in Table 4 for the 2-amino pentanoic acid ("Ape"; side chain = (Cα)-CH$_2$-CH$_2$-CH$_3$), Ile, and Leu side chains. These energies are given relative to the lowest energy for each r_1 rotamer, and therefore represent the interactions responsible for the parameters of the *conditional* backbone-independent rotamer library described above.

The r_1 and r_2 rotamers of Lys, Arg, Met, Glu, Gln, Leu, and Ile are exactly analogous to the g^+, t, and g^- rotamers of butane and pentane, because, for these side chains, the χ_1 and χ_2 dihedral rotations are about sp^3–sp^3 carbon–carbon bonds (χ_1 = Cα-Cβ and χ_2 = Cβ-Cγ). For the aromatic side chains and Asn and Asp, the γ heavy atom has an sp^2 hybridization state, and therefore the χ_2 rotation is not described by the g^+, t, g^- rotameric states.

The energy curves of the peptide fragment of 2-amino pentanoic acid are given in Figure 8A as a function of χ_2 for each of the three r_1 rotamers. Four of the nine local minima have syn-pentane in-

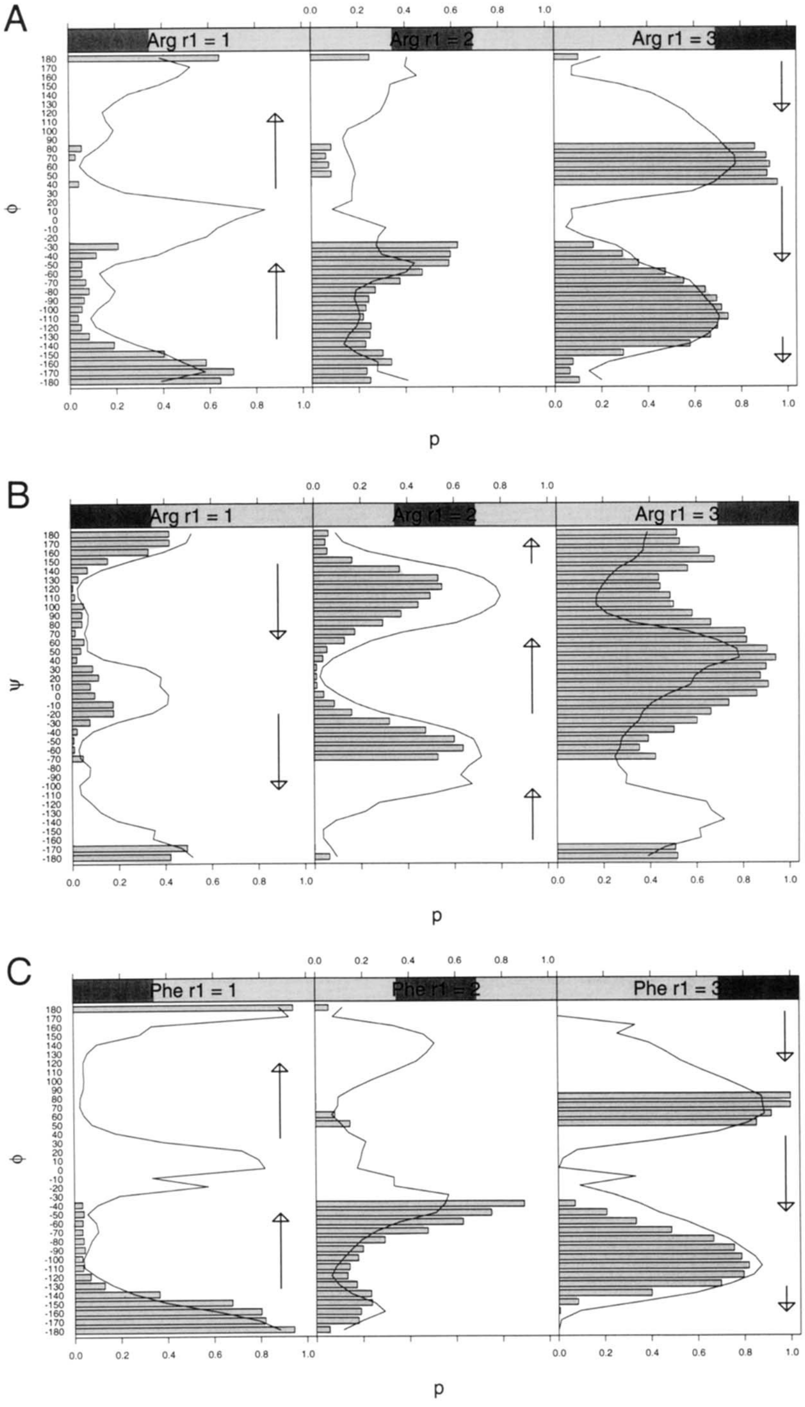

Fig. 3. Data calculated with the nonperiodic weighting function (solid bars) for $y_a^{non\text{-}per}$ and $y_b^{non\text{-}per}$ and the prior density function from W^{per} (lines) for (**A**) Arg, ϕ; (**B**) Arg, ψ; (**C**) Phe, ϕ; (**D**) Phe, ψ; (**E**) Asp, ϕ; (**F**) Asp, ψ; (**G**) Val, ϕ; (**H**) Val, ψ. Regions of ϕ or ψ where interactions between backbone atoms and Cγ are likely to lower the populations of certain rotamers are marked with arrows. Arrows point from values of the dihedral connecting the backbone atom to Cβ from −90° to 0° or from +90° to 0°. See Table 5. (*Figure continues on following pages.*)

teractions: when N_i and Cδ (χ_1,χ_2) are connected by $\{g^+,g^-\}$ or $\{g^-,g^+\}$ dihedrals, and when C_i and Cδ ($\chi_1-120°,\chi_2$) are connected by $\{g^+,g^-\}$ or $\{g^-,g^+\}$ dihedrals. These four are evident in Figure 8A at the $\{g^+,g^+\}$, $\{g^+,g^-\}$, $\{t,g^-\}$, and $\{g^-,g^+\}$ positions. Three of the remaining five rotamers have gauche interactions ($\{g^+,t\}$, $\{t,g^+\}$, and $\{g^-,g^-\}$), raising the energy by 0.5–0.9 kcal/mol over the global minima, $\{t,t\}$ and $\{g^-,t\}$. These interactions are listed in Table 4 and rotamer pairs with syn-pentane interactions are indicated in the experimental data listed in Table 2, all with very low probabilities and with skewed χ angles.

Ile resembles the Ape side chain, except that the $r_1 = g^-$ rotamer is lower in energy than the other two rotamers because of

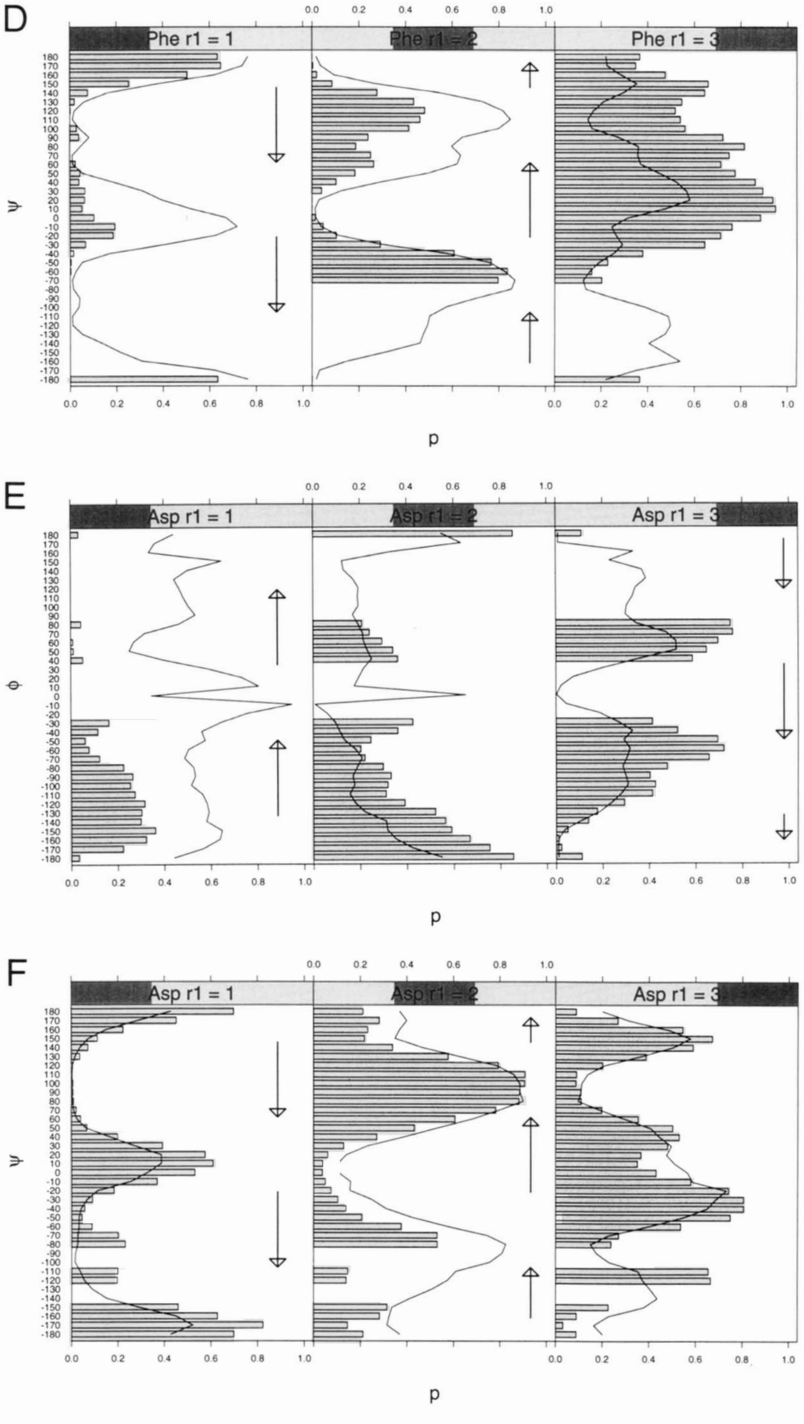

Fig. 3. *Continues.*

gauche interactions with backbone N$_i$ and C$_i$ (equivalent to Val *t* rotamer). Leu has δ atoms at χ_2 and $\chi_2 + 120°$, and so there is a total of eight syn-pentane interactions. These occur for all three $r_1 = g^+$ rotamers, and so Leu is very unlikely to be in a g^+ rotamer. The conditional probabilities for the $r_1 = g^+$ rotamer of Leu in Table 2 shows that the $\{g^+, g^-\}$ rotamer has much lower probability than the other two g^+, g^+ and g^+, t. The reason is evident from Table 4, because the $\{g^+, g^-\}$ combination has two syn-pentane interactions with the backbone, whereas the other two $r_1 = g^+$ rotamers have only one.

In Figure 8B, the energies of Phe rotamers demonstrate that the r_1 rotamer affects the position of the local χ_2 minimum. Assuming the sp^3–sp^2 rotation about Cα-Cβ has minima at +90° and −90°, there is some deviation in the *t* and g^- r_1 rotamers. These are caused by interactions of the Cδ atoms with backbone N$_i$ and C$_i$: when r_1 is g^-, a syn-pentane interaction when χ_2 is below 90° pushes the minimum to $\chi_2 = +100°$; when r_1 is *t*, a syn-pentane interaction when χ_2 is above 90° pushes the minimum down to $\chi_2 = +78°$. g^+ rotamers are always higher than g^- and *t* rotamers

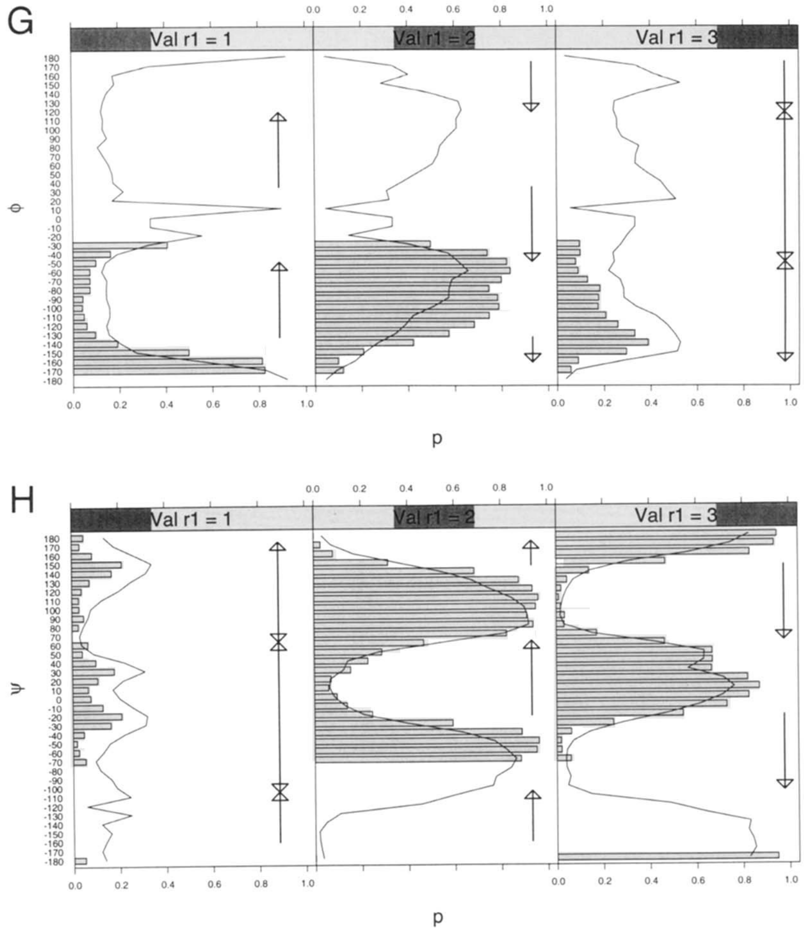

Fig. 3. *Continued.*

by more than 1.5 kcal/mol because of syn-pentane interactions between $C\delta_1$ and $C\delta_2$ and N_i and C_i.

We note that typical molecular mechanics calculations for $p(r_1, r_2, r_3, r_4)$ (Nayeem & Scheraga, 1994) resemble our $\theta_{jkl|i}$, and not θ_{ijkl}. The traditional backbone-independent rotamer library parameters, $p(r_1, r_2, r_3, r_4)$, are determined both by the relative energies of the three rotamers in each backbone conformation, and by the populations of side chains in different backbone conformations (i.e., the Ramachandran map distribution). They can be calculated by integrating the product of the conditional backbone-independent probabilities and the backbone-dependent r_1 probabilities over the whole Ramachandran map,

$$E(\theta_{ijkl}) \equiv p(r_1 = i, r_2 = i, r_3 = i, r_4 = i)$$
$$= \int p(\theta_{jkl|i}) p(\theta_i|\phi,\psi) p(\phi,\psi)\, d\phi\, d\psi. \tag{4}$$

Conformational analysis of backbone-conformation-dependent interactions

Rotamer populations are affected not only by backbone N_i and C_i of residue i whose positions are independent of ϕ and ψ, but also by the positions of other backbone atoms, especially C_{i-1}, O_i, and N_{i+1}, whose positions are dependent on ϕ and ψ. These three atoms are all connected to γ heavy atoms by two sp^3 hybridized atoms ($C\beta$ and $C\alpha$) and one sp^2 hybridized atom (backbone N_i or C_i). There are therefore two dihedral degrees of freedom separating the γ heavy atoms with the backbone atoms C_{i-1}, H_i, O_i, and N_{i+1} (see Fig. 7B for the Newman projection of the side-chain with the dipeptide backbone added). We include H_i and its hydrogen bond acceptor, because there are apparent shifts in rotamer populations due to this interaction. Hence, there is the possibility of syn-pentane interactions between any of these atoms and side-chain γ atoms when the connecting dihedrals occur in $\{g^+, g^-\}$ or $\{g^-, g^+\}$ combinations. The interactions are of smaller magnitude, because bond angles at the sp^2 hybridized atoms are 120° instead of 109.5°. They are nevertheless of sufficient magnitude to alter the rotamer distributions significantly. The syn-pentane interactions that occur between γ heavy atoms and backbone atoms dependent on ϕ and ψ are listed in Table 5, along with the probable ranges of interaction. As with pentane, the range of interaction is about 90°.

We calculated the energies of the χ_1 rotamers of several side chains to investigate how well the CHARMM potential describes the experimental distributions of the backbone-dependent rotamer library. This was done by minimizing the energy of the dipeptide of each residue type (N-acetyl-Xxx-N′-methylamide) with ϕ and ψ constrained to values in 10° increments. The probability surfaces of the three rotamers of 2-amino butanoic acid (Abu) and valine as

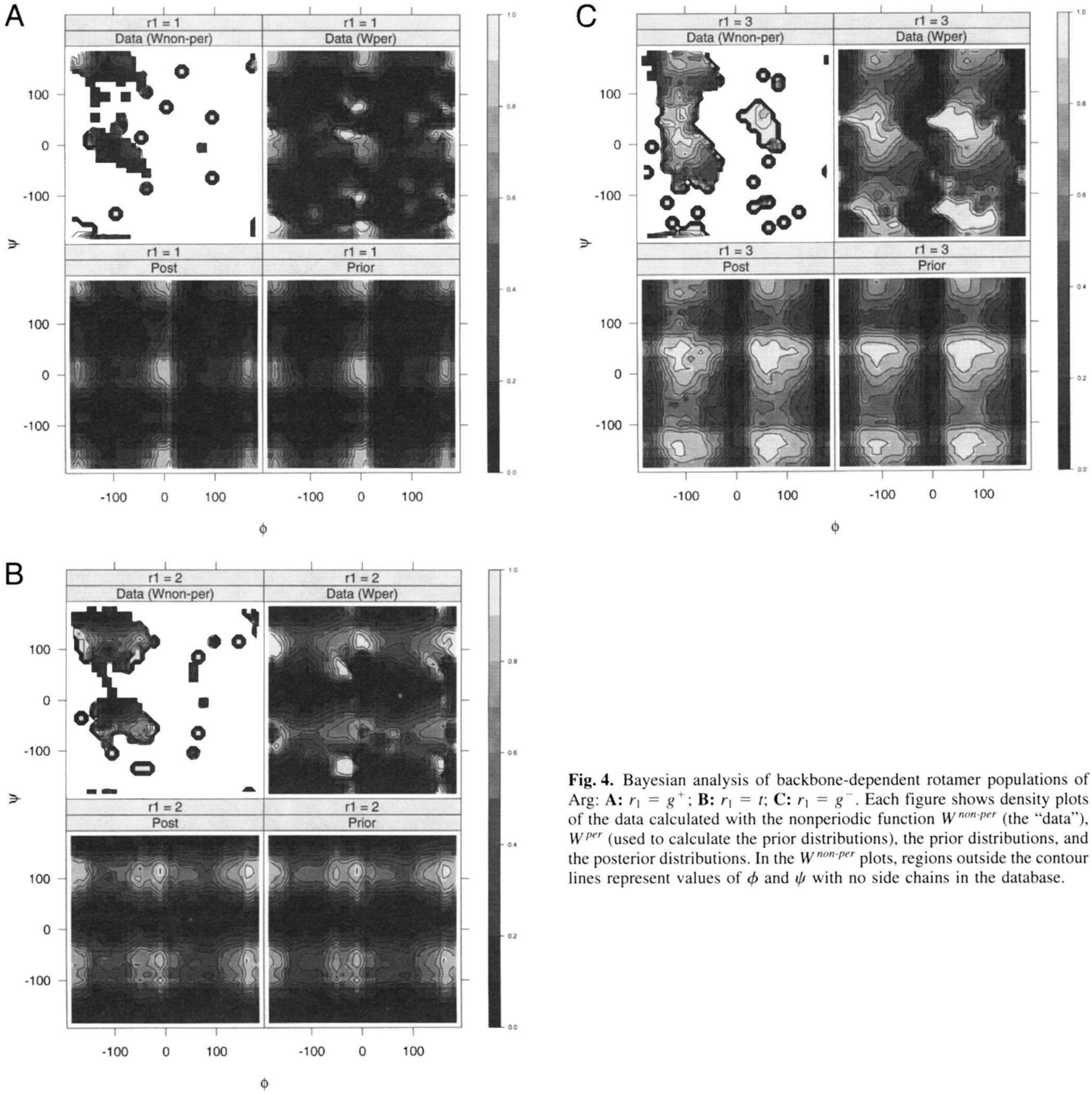

Fig. 4. Bayesian analysis of backbone-dependent rotamer populations of Arg: **A:** $r_1 = g^+$; **B:** $r_1 = t$; **C:** $r_1 = g^-$. Each figure shows density plots of the data calculated with the nonperiodic function $W^{non\text{-}per}$ (the "data"), W^{per} (used to calculate the prior distributions), the prior distributions, and the posterior distributions. In the $W^{non\text{-}per}$ plots, regions outside the contour lines represent values of ϕ and ψ with no side chains in the database.

a function of ϕ and ψ are shown in Figure 9A and B, respectively. To produce these figures, we assumed $kT = 1$.

When $r_1 = g^+$, we expect a syn-pentane interaction when $\phi = 60°$ (Cγ and C$_{i-1}$) and $-120°$ (Cγ and O$\cdots$HN$_i$) and when $\psi = -60°$ (Cγ and N$_{i+1}$) and 120° (Cγ and O). These are shown clearly as the vertical and horizontal darker regions in the first panel of Figure 9A for the Abu side-chain. These are also evident as troughs in the experimental data in the $r_1 = 1$ panels of Figure 3 (marked by arrows). These interactions also appear as the dark regions in the $r_1 = 1$ panels of Figure 4 and Figure 5. The Cγ/C$_{i-1}$ interaction is the largest, approximately 3.2 kcal/mol, whereas the Cγ/N$_{i+1}$ and Cγ/O$_i$ interactions are 1.6 and 1.2 kcal/mol, respectively. In the second panel of Figure 9A, the interactions of t rotamer Cγ with N$_{i+1}$ when $\psi = \pm180°$ and Cγ with O$_i$ when $\psi = 0°$ are also evident. These probability minima are also evident experimentally in the $r_1 = 2$ panels of Figures 3, 4, and 5. Again, these interactions are marked by arrows in Figure 3. The third panel of Figure 9A confirms the Cγ/C$_{i-1}$ interaction when $\phi = \pm180°$ and near 0° for the g^- r_1 rotamer, as shown in the data in Figures 3, 4, and 5. It is worth noting the weak ψ dependence in the g^- panels, reflecting changes in the energies of the other two rotamers as a function of ψ. It is clear from the three maps that the $\chi_1 = -60°$ conformation of Abu is the most favored in much of the heavily occupied portions of the Ramachandran map (i.e., $-120° \leq \phi \leq -40°$; $-80° \leq \psi \leq -30°$ and $100° \leq \psi \leq 150°$). $r_1 = t$ has steric interactions with backbone atoms in extended backbone conformations and

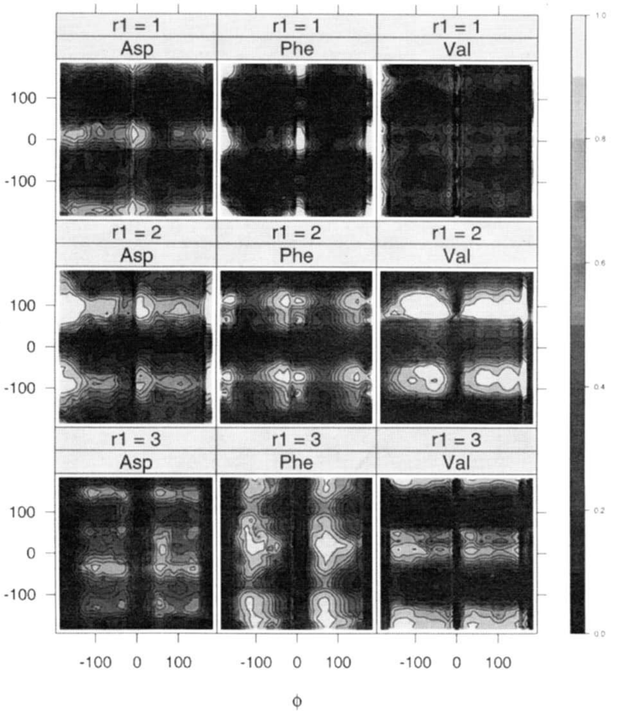

Fig. 5. Average backbone-dependent posterior parameters for Asp, Phe, and Val. The averages are given by the expectation values of Equation 14.

near $\psi = 0°$, whereas $r_1 = 60°$ has steric conflicts in β-sheets as well as α-helices.

In Figure 9B, the CHARMM probability and energy surfaces for Val are shown. The alternation between t and g^- rotamers with variation in ψ is due to interactions with each γ carbon with backbone N_{i+1} and O_i, which result in the troughs at $\psi = 0°$ and 180° for the second panel (t) and at $\psi = -60°$ and 120° in the third panel (g^-). Both of these rotamers have steric interactions with C_{i-1} at $\phi = -180°$, and so the g^+ rotamer in the first panel has high density in this region. These figures can be compared with the experimental probabilities from the Bayesian analysis for Val in Figure 3G and H, and Figure 5.

Finally, in Table 6, we show the prediction rates for the backbone-dependent rotamer library. The data in this table were obtained by predicting the side-chain conformation for each side chain in the database based on the identity of the r_1 rotamer with the largest probability in the posterior distribution, given the values of ϕ and ψ for that side chain. No optimization or removal of steric clashes was performed. The library is able to predict 73% of χ_1's of side chains that have structures within 40° of one of the canonical sp^3–sp^3 rotamers (+60°, 180°, −60°). From the steric analysis and the CHARMM calculations, it seems that most of these are sterically the lowest energy minima (i.e., ignoring electrostatic interactions in polar side chains) given the local backbone dihedral angles ϕ and ψ. Another 22% of side chains adopt the next most likely local minimum energy structure, and the final 5% the highest energy, least likely local minimum. Close to 4% of all side chains are not within 40° of one of the g^+, g^-, and t rotamer conformations. It is likely that most of these are either an average of two occupied rotamers, or an error of interpretation of electron density, given the high energy price paid (3–6 kcal/mol) for these strained conformations and that proteins are only marginally stable ($\Delta G_{folding}$ = 5–20 kcal/mol). It would seem that a protein would not be able to afford more than one or two such side-chain conformations, and it would therefore be useful to scrutinize the electron density carefully in such cases to determine if there are interactions that explain the high-energy conformation or whether a more plausible conformation can be found (Schrauber et al., 1993).

These results can be considered minimum values for any side-chain prediction method for the self-backbone prediction test, because they were obtained with no computational effort beyond looking up the best rotamer in the database. We have recently used the backbone-dependent rotamer library described in this paper in a prediction program that removes steric clashes after the backbone-dependent rotamer library prediction (Bower et al., 1997). Removing steric clashes in a set of 299 self-backbone side-chain prediction tests raises the prediction rate from 73% to 78%. We compared X-ray structures of identical proteins in different crystal space groups and found that only 80% of side chains retain their r_1 rotamers in different crystal structures. It is likely that a large fraction of the remaining 20% are in two different r_1 rotamers, one of which is the lowest energy rotamer for the local backbone conformation.

Discussion

Understanding the factors that determine protein side-chain conformation and its dynamics is important in a number of areas. First, the fact that the backbone has such a strong effect on side-chain conformation distributions is likely to be exploited by proteins as they fold. This effect may also operate as proteins evolve if mutations of residues with unfavorable steric interactions to residues without these interactions increase the stability of the folded protein. This topic has been studied extensively using site-directed mutagenesis to probe the structural and thermodynamic implications of a side chain's interaction with its environment. Mutations that increase or decrease local backbone–side-chain steric interactions might be expected to have some effect on the stability of the folded structure, because these interactions are presumably absent in the unfolded state.

Second, homology modeling of proteins for the purposes of drug design depends mostly upon an accurate representation of the side-chain conformations in the target structure. It has proved very difficult to model the insertion and deletion of loops in proteins, but, in cases where these insertions are far from the binding site of an enzyme substrate or allosteric effector, it is only the binding site side-chain conformations that must be constructed carefully. Because proteins favor their dipeptide low-energy rotamers in their folded states, the backbone-dependent rotamer library is especially helpful in predicting side-chain conformations in homology modeling situations where no information is present in the template structure (e.g., Ala → Val; Pro → Phe). It is notable that many current side-chain placement methods use rotamer libraries dependent on the local structural environment of side chains (Moult & James, 1986; Levitt, 1992; Bower et al., 1997), including those used in several popular molecular modeling packages, e.g., QUANTA (Dunbrack & Karplus, 1993), SYBYL (Schrauber et al., 1993), and WHAT-IF (Chinea et al., 1995).

We have set up a website of the conformational analysis of protein side chains (Dunbrack, 1997) as described in this paper. The website includes both backbone-independent and backbone-

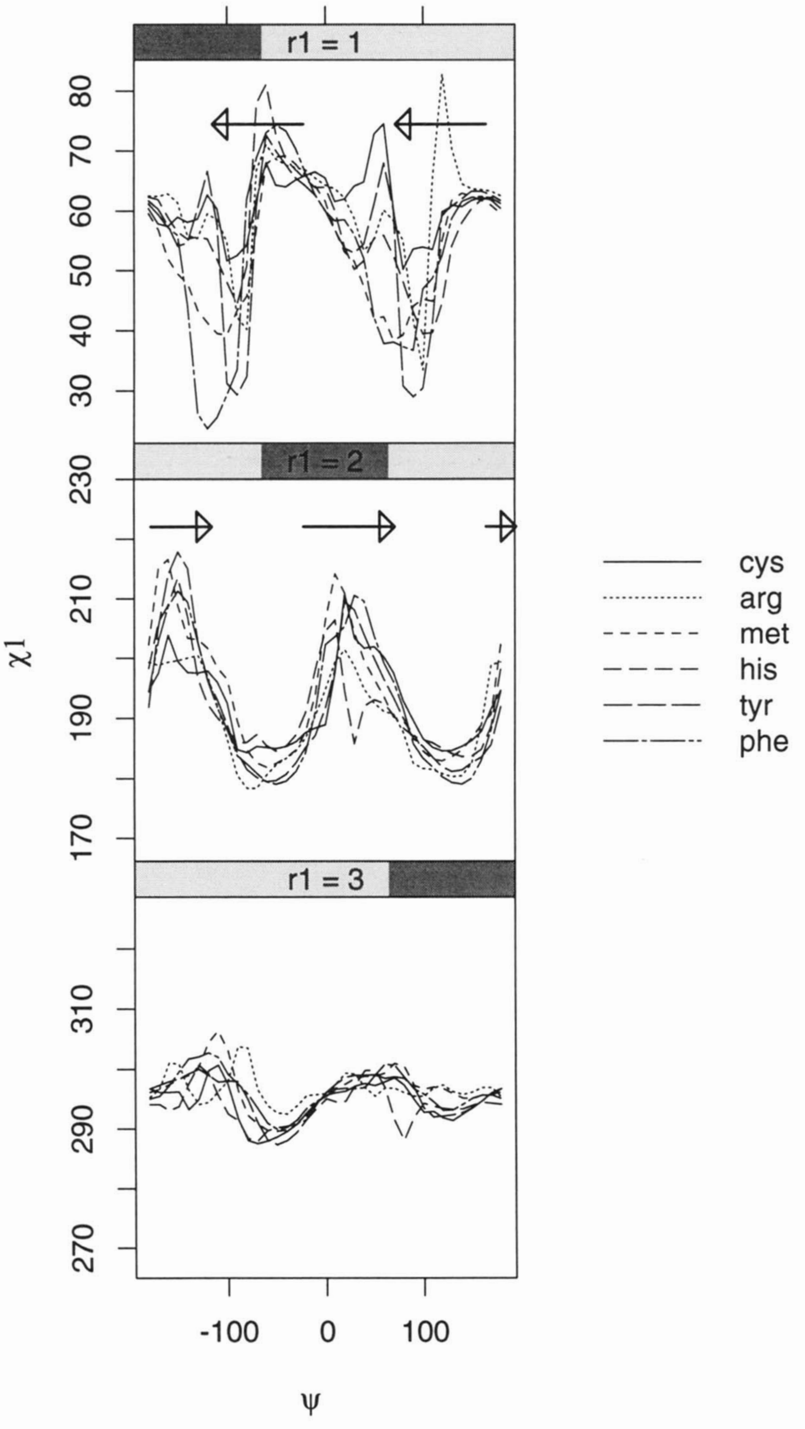

Fig. 6. Average χ angle ψ-dependent prior distributions for aromatic side chains The values given for each angle are deviations from the canonical values of $+60°$, $180°$, and $-60°$, viz. $\chi_1 -60°$, $\chi_1 -180°$, and $\chi_1 +60°$ for the $r_1 = 1,2,3$ rotamers. Arrows indicate ranges of syn-pentane interactions between side-chain Cγ and backbone atoms.

dependent analysis as well as "raw" data counts and the Bayesian posterior distributions described here. We have shown graphical representations of our results for several side chains in this paper. On the website we make available the graphical representations for all side-chain types. As the PDB continues to grow, we will update the results presented here periodically and place them on the website.

Although many of the steric effects in backbone-independent and dependent side-chain conformations described in this paper have been studied previously in protein structures (Chandrasekaran & Ramachandran, 1970; Cody et al., 1973; Janin et al., 1978; Bhat et al., 1979; Benedetti et al., 1983; James & Sielecki, 1983; McGregor et al., 1987; Ponder & Richards, 1987; Sutcliffe et al.,1987; Tuffery et al., 1991) and by energy calculations (Pon-

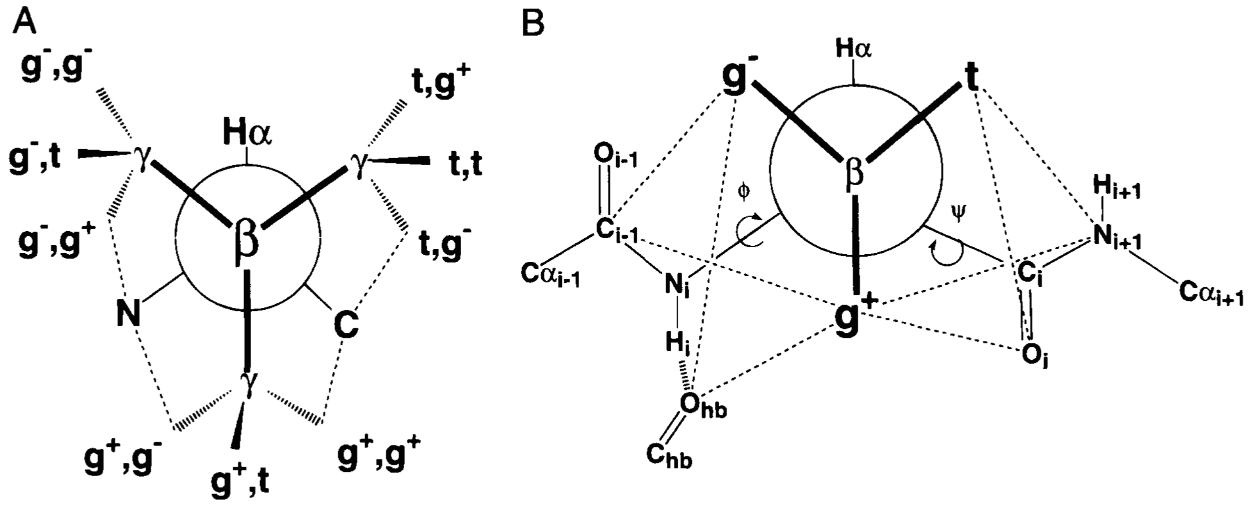

Fig. 7. Newman projections for conformational analysis of protein side chains. **A:** Backbone-independent interactions. **B:** Backbone-dependent interactions. Dotted lines mark possible syn-pentane interactions.

nuswamy & Sasisekharan, 1971; Sasisekharan & Ponnuswamy, 1971; Lewis et al., 1973; Janin et al., 1978; Zimmerman & Scheraga, 1978; Gelin & Karplus, 1979; Benedetti et al., 1983), we believe the present analysis is more thorough and consistent in its approach. In particular, the conformational analysis of gauche and syn-pentane interactions has provided a simple organizing principle for explaining and predicting the effects of backbone conformation on side-chain conformation that previously has not been used extensively on peptides. We hope the library will be useful in a number of applications as well as furthering our understanding of the determinants of protein conformation and the process of protein folding.

Materials and methods

Experimental data

To update the backbone-dependent rotamer library, we obtained a list of protein chains with 2.0 Å resolution or better and less than 50% sequence identity with other chains in the list from the OBSTRUCT server of Heringa et al. (1992). The present library is based on 518 chains, which is four times larger than the original library. The list of proteins used is available on the Backbone-Dependent Rotamer Library Website (Dunbrack, 1997).

Rotamers r_1, r_2, r_3, and r_4 were defined according to the limits on χ angles listed in Table 1 for each amino acid type. Residues with missing backbone atoms necessary to determine ϕ and ψ (i.e., the first and last residue of a chain) or with missing side-chain atoms were discarded. As in the previous work, we assumed that, in crystal structures, the χ_2 orientations of the His and Asn side chains and the χ_3 orientation of Gln are not uniquely determined under a rotation of 180°.

For the raw backbone-independent library, we used the limits in Table 1 to count the number of side chains with $r_1,r_2,r_3,r_4 = i,j,k,l$ for all side-chain types. These numbers, y_{ijkl}, can be used to calculate various raw conditional probabilities,

$$p(r_2 = j|r_1 = i) \equiv \beta_{j|i} = \frac{\sum_{k,l} y_{ijkl}}{\sum_{j,k,l} y_{ijkl}}$$

$$p(r_3 = k|r_2 = j) \equiv \beta_{k|j} = \frac{\sum_{i,l} y_{ijkl}}{\sum_{i,k,l} y_{ijkl}}$$

$$p(r_4 = l|r_3 = k) \equiv \beta_{l|k} = \frac{\sum_{i,j} y_{ijkl}}{\sum_{i,j,l} y_{ijkl}} \tag{5}$$

which will be used in the Bayesian analysis.

To be consistent with our previous work (Dunbrack & Karplus, 1993, 1994), we first calculated a "raw" count of each r_1 in 20° by 20° ϕ,ψ bins centered 10° apart (−180°, −170°, ..., 0°, ..., 160°, 170°) in ϕ and ψ. Each sidechain is counted four times in such a procedure. Mathematically, we can define a function W^{box} for a bin centered at $\{\phi_a,\psi_b\}$ that is applied to all side chains with rotamer $r_1 = i$, such that

$$y^{box}_{i|ab} = \sum_{\substack{\text{sidechains } m \\ \text{with } r_1=i}} W^{box}_{ab}(\Delta\phi_m,\Delta\psi_m), \tag{6}$$

where $\Delta\phi_m = \phi_m - \phi_a$ and $\Delta\psi_m = \psi_m - \psi_b$. $W^{box} = 1$ if both $\Delta\phi_m$ and $\Delta\psi_m$ are less than 10°; $W^{box} = 0$ otherwise. The function W^{box} is shown in Figure 2 for $\{\phi_a,\psi_b\} = \{-60°, -60°\}$. The corresponding probabilities are denoted $\beta^{box}_{i|ab} = y^{box}_{i|ab}/\sum_{i'} y^{box}_{i'|ab}$.

To provide estimates for the underpopulated and unpopulated regions of the map, we used a Gaussian weighted periodic function to count rotamers. Side chains were counted for each r_1 rotamer type with the following weight function in ϕ,ψ bins centered 10° apart (−180°, −170°, ..., 0°, ..., 170°):

Table 4. *Conformational analysis of backbone-independent interactions: Single Cγ, single Cδ (Met, Glu, Gln, Arg, Lys)*

				Met, Glu, Gln, Arg, Lys r_1r_2 rotamers						
N-Cα-Cβ-Cγ	χ_1	g^+	g^+	g^+	t	t	t	g^-	g^-	g^-
C-Cα-Cβ-Cγ	$\chi_1-120°$	g^-	g^-	g^-	g^+	g^+	g^+	t	t	t
Cα-Cβ-Cγ$^-$Cδ	χ_2	g^+	t	g^-	g^+	t	g^-	g^+	t	g^-
N-Cα-Cβ-Cγ$^-$Cδ[a]	χ_1,χ_2			p				p		
C-Cα-Cβ-Cγ$^-$Cδ	$\chi_1-120°,\chi_2$	p					p			
E[b]		p + 3g	2g	p + 3g	2g	1g	p + 2g	p + 2g	1g	2g
ΔE		p + g		p + g	g		p + g	p + g		g
CHARMM ΔE[c]		2.66	0.00	2.44	0.42	0.00	2.28	1.91	0.00	0.42
Arg[d]	$p(r_2\|r_1)$	5.1	91.9	3.1	15.7	80.1	4.2	4.2	75.7	20.2
Met	$p(r_2\|r_1)$	6.3	89.2	4.6	30.4	62.5	7.1	1.7	58.7	39.6
Gln	$p(r_2\|r_1)$	8.6	73.5	17.9	33.4	60.6	6.0	7.1	62.9	30.0
Glu	$p(r_2\|r_1)$	5.3	65.3	29.4	21.7	72.1	6.2	13.4	60.6	26.1
Lys	$p(r_2\|r_1)$	5.9	88.5	5.7	18.2	76.5	5.4	6.5	69.3	24.2
				Ile r_1, r_2 rotamers						
N-Cα-Cβ-Cγ1	χ_1	g^+	g^+	g^+	t	t	t	g^-	g^-	g^-
N-Cα-Cβ-Cγ2	$\chi_1-120°$	g^-	g^-	g^-	g^+	g^+	g^+	t	t	t
C-Cα-Cβ-Cγ1	$\chi_1-120°$	g^-	g^-	g^-	g^+	g^+	g^+	t	t	t
C-Cα-Cβ-Cγ2	$\chi_1+120°$	t	t	t	g^-	g^-	g^-	g^+	g^+	g^+
Cα-Cβ-Cγ1-Cδ	χ_2	g^+	t	g^-	g^+	t	g^-	g^+	t	g^-
N-Cα-Cβ-Cγ1-Cδ	χ_1,χ_2			p				p		
C-Cα-Cβ-Cγ1-Cδ	$\chi_1-120°,\chi_2$	p					p			
E		p + 4g	3g	p + 4g	4g	3g	p + 4g	p + 3g	2g	3g
ΔE		p + g		p + g	g		p + g	p + g		g
CHARMM ΔE		1.93	0.00	2.51	0.00	0.23	2.42	1.95	0.00	0.18
Ile	$p(r_2\|r_1)$	9.4	89.4	1.2	31.2	66.4	2.6	3.5	76.2	20.4
				Leu r_1, r_2 rotamers						
N-Cα-Cβ-Cγ	χ_1	g^+	g^+	g^+	t	t	t	g^-	g^-	g^-
C-Cα-Cβ-Cγ	$\chi_1-120°$	g^-	g^-	g^-	g^+	g^+	g^+	t	t	t
Cα-Cβ-Cγ$^-$Cδ1	χ_2	g^+	t	g^-	g^+	t	g^-	g^+	t	g^-
Cα-Cβ-Cγ$^-$Cδ2	$\chi_2+120°$	t	g^-	g^+	t	g^-	g^+	t	g^-	g^+
N-Cα-Cβ-Cγ$^-$Cδ1	χ_1,χ_2			p				p		
N-Cα-Cβ-Cγ$^-$Cδ2	$\chi_1,\chi_2+120°$		p							p
C-Cα-Cβ-Cγ$^-$Cδ1	$\chi_1-120°,\chi_2$	p					p			
C-Cα-Cβ-Cγ$^-$Cδ2	$\chi_1-120°,\chi_2+120°$			p		p				
E		p + 3g	p + 3g	2p + 4g	2g	p + 2g	p + 3g	p + 2g	2g	p + 3g
ΔE				p + g		p	p + g	p		p + g
CHARMM ΔE		0.25	0.00	2.77	0.00	1.94	2.09	1.78	0.00	1.97
Leu	$p(r_2\|r_1)$	55.9	41.3	2.8	84.4	12.2	3.3	13.3	83.7	3.1

[a] "p" denotes a syn-pentane interaction between the atom 1 and 5 of the listed in the first column.

[b] As a rough estimate, one gauche ("g") interaction is ~ 0.9 kcal/mol; p ~ 1.5 kcal/mol, based on pentane ab initio calculations (Wiberg & Murcko, 1988).

[c] CHARMM energies are given relative to the lowest r_2 rotamer for each r_1 rotamer.

[d] Experimental data for the conditional probabilities, $p(r_2|r_1)$.

$$y_{i|ab}^{per} = \sum_{\substack{\text{sidechains } m \\ \text{with } r_1=i}} \exp(-k_1 \sin^2(\Delta\phi_m) - k_1 \sin^2(\Delta\psi_m)) \times \exp\left(-\left(\frac{\Delta\phi_m}{k_2}\right)^2 - \left(\frac{\Delta\psi_m}{k_2}\right)^2\right). \tag{7}$$

k_1 is a constant that determines the width of the function about $\{\phi_a,\psi_b\}$ and k_2 determines the ratio of the peak heights at $\{\phi_a,\psi_b\}$, $\{\phi_a \pm 180°,\psi_b\}$, $\{\phi_a,\psi_b \pm 180°\}$, and $\{\phi_a \pm 180°, \psi_b \pm 180°\}$. We used values of $k_1 = 20$ and $k_2 = 2.0$ radians. This function is also plotted in Figure 2 for $\{\phi_a,\psi_b\} = \{-60°,-60°\}$. The corre-

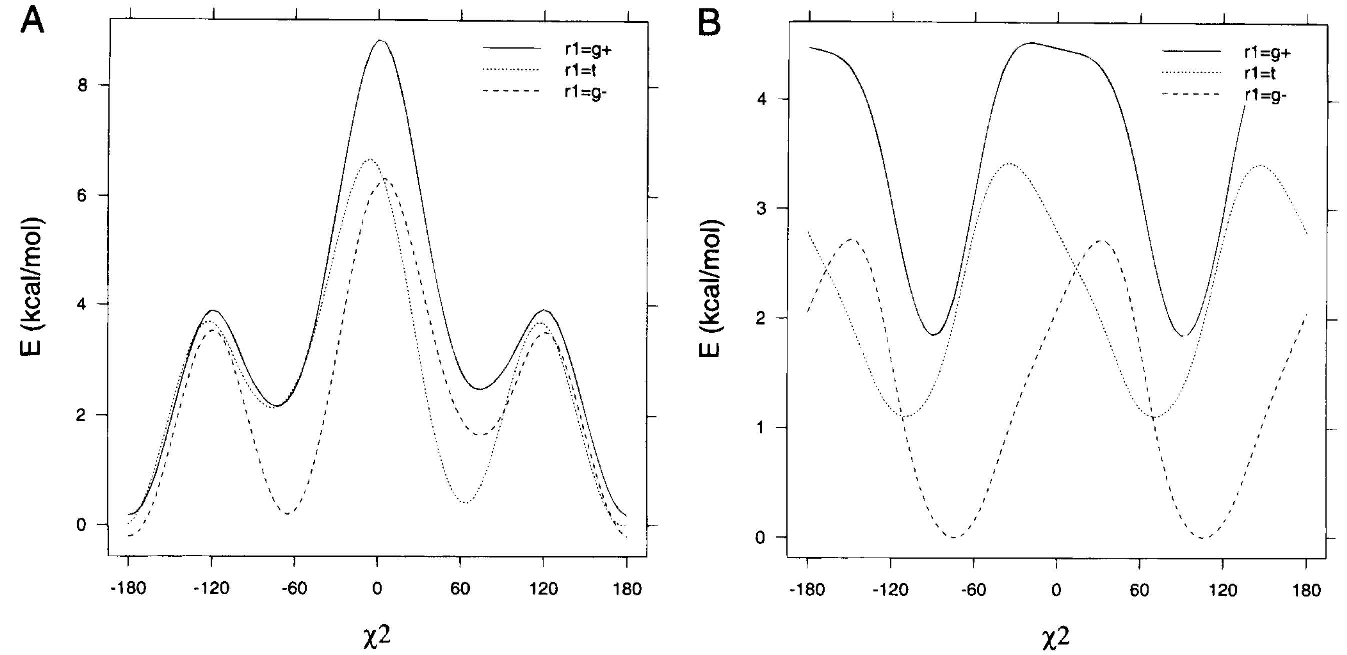

Fig. 8. CHARMM calculations of backbone-conformation independent backbone–side-chain interactions. **A**: Ape. **B**: Phe.

sponding probabilities, $\beta^{per}_{i|ab}$ were calculated by normalizing Equation 7 for each ϕ_a,ψ_b pair, $\beta^{per}_{i|ab} = y^{per}_{i|ab}/\sum_{i'} y^{per}_{i'|ab}$.

We used a third function, $W^{non\text{-}per}$ to count rotamers for the data used in the Bayesian analysis. This function is a product of W^{per} and W^{box}. It counts side chains with an approximately Gaussian weight in the neighborhood of each ϕ,ψ point. But because of the box function W^{box}, side chains with $\Delta\phi$ or $\Delta\psi$ greater than 10° are not counted by $W^{non\text{-}per}$ (see Fig. 2).

Bayesian statistical analysis

Rotamer statistics

We derive a library used for predictions of side-chain rotamers from the PDB by a Bayesian statistics analysis (Gelman et al., 1995). In this section, we describe Bayesian statistical analysis briefly; further details for the rotamer library are described below. We follow the notation of Gelman et al. (1995). We are considering

Table 5. *Conformational analysis of backbone-dependent interactions*

Atoms					Connecting dihedrals		Interaction: −90° → 0°, g^+		Interaction: 0° → 90°, g^-	
1	2	3	4	5	1-2-3-4	2-3-4-5	1-2-3-4 = −90° → 0°	2-3-4-5 = g^+	1-2-3-4 = 0° → 90°	2-3-4-5 = g^-
All Cγ, Cys Sγ, Ser Oγ, Val Cγ1, Ile Cγ1, Thr Oγ1										
C_{i-1}	N	Cα	Cβ	Cγ	$\phi-120°$	χ_1	$\phi =$ 30° → 120°	$r_1 = g^+$	$\phi =$ −240° → −150°	$r_1 = g^-$
O=C···HN	N	Cα	Cβ	Cγ	$\phi+60°$	χ_1	$\phi =$ −150° → −60°	$r_1 = g^+$	$\phi =$ −60° → 30°	$r_1 = g^-$
N_{i+1}	C	Cα	Cβ	Cγ	$\psi+120°$	χ_1 −120°	$\psi =$ 150° → 240°	$r_1 = t$	$\psi =$ −120° → −30°	$r_1 = g^+$
O	C	Cα	Cβ	Cγ	$\psi-60°$	χ_1 −120°	$\psi =$ −30° → 60°	$r_1 = t$	$\psi =$ 60° → 150°	$r_1 = g^+$
Val Cγ2										
C_{i-1}	N	Cα	Cβ	Cγ2	$\phi-120°$	$\chi_1+120°$	$\phi =$ 30° → 120°	$r_1 = g^-$	$\phi =$ −240° → −150°	$r_1 = t$
O=C···HN	N	Cα	Cβ	Cγ2	$\phi+60°$	$\chi_1+120°$	$\phi =$ −150° → −60°	$r_1 = g^-$	$\phi =$ −60° → 30°	$r_1 = t$
N_{i+1}	C	Cα	Cβ	Cγ2	$\psi+120°$	χ_1	$\psi =$ 150° → 240°	$r_1 = g^+$	$\psi =$ −120° → −30°	$r_1 = g^-$
O	C	Cα	Cβ	Cγ2	$\psi-60°$	χ_1	$\psi =$ −30° → 60°	$r_1 = g^+$	$\psi =$ 60° → 150°	$r_1 = g^-$
Ile Cγ2 and Thr Cγ2										
C_{i-1}	N	Cα	Cβ	Cγ2	$\phi-120°$	$\chi_1-120°$	$\phi =$ 30° → 120°	$r_1 = t$	$\phi =$ −240° → −150°	$r_1 = g^+$
O=C···HN	N	Cα	Cβ	Cγ2	$\phi+60°$	$\chi_1-120°$	$\phi =$ −150° → −60°	$r_1 = t$	$\phi =$ −60° → 30°	$r_1 = g^+$
N_{i+1}	C	Cα	Cβ	Cγ2	$\psi+120°$	$\chi_1+120°$	$\psi =$ 150° → 240°	$r_1 = g^-$	$\psi =$ −120° →. −30°	$r_1 = t$
O	C	Cα	Cβ	Cγ2	$\psi-60°$	$\chi_1+120°$	$\psi =$ −30° → 60°	$r_1 = g^-$	$\psi =$ 60° → 150°	$r_1 = t$

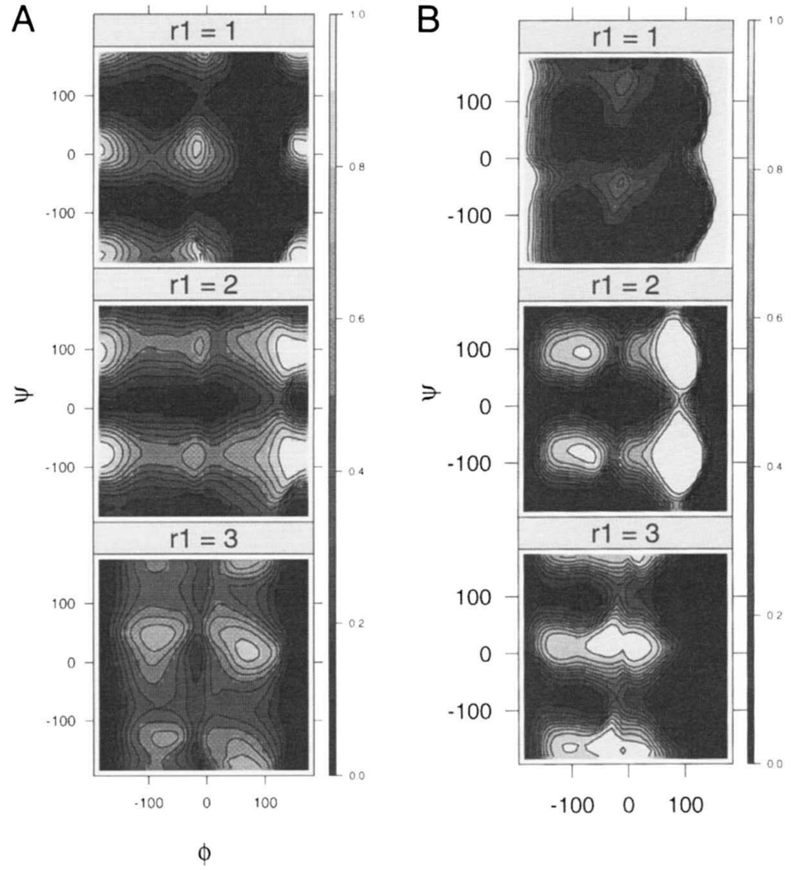

Fig. 9. CHARMM calculations of backbone-conformation dependent backbone–side-chain interactions. **A:** Abu probabilities; **B:** Val probabilities. Probabilities were calculated from Equation 28, assuming $kT = 1.0$.

a population of side chains in different states or rotamers. We refer to the probabilities of the different rotamers as θ_i, so for the three r_1 rotamers we have θ_1, θ_2, and θ_3. These are *superpopulation* parameters; that is, the probabilities of the three different rotamers in a hypothetical infinitely large Protein Data Bank. The observed data, y_1, y_2, and y_3, are the numbers of rotamers observed in the finite experimental data of the PDB. Unless the total of $y_1 + y_2 + y_3 = N$ is large (>50), the observed populations y_i/N may differ somewhat from the θ_i. Bayesian analysis provides a method for calculating a probability distribution of the *parameters* θ based on the *observed data* y, where θ and y are vectors with m components ($m = 3$ in this example).

We start with the *joint probability distribution* for θ and y written with Bayes' rule as the product of a *prior distribution* $p(\theta)$ and the *likelihood* or *sampling distribution* $p(y|\theta)$,

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{8}$$

The prior distribution contains any information we may have about the expected value of θ in the form of a probability density over the acceptable values of θ. We are looking for an expression for $p(\theta|y)$, which is the probability distribution of the parameters of interest, θ, conditioned on the observed experimental data, y. Using Bayes' rule again, such that, $p(\theta,y) = p(y)p(\theta|y)$, we find,

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}. \tag{9}$$

Because y is fixed, we can write the *posterior distribution* $p(\theta|y)$ as proportional to the *unnormalized posterior density*,

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{10}$$

So if we have some guess for the prior distribution $p(\theta)$ and a model for the likelihood function $p(y|\theta)$ for the probability of the data given values for the parameters, we can derive a posterior probability distribution for the parameters, $p(\theta|y)$.

There is a distinction in Bayesian statistics between *informative* and *noninformative* prior distributions. Noninformative prior distributions are usually flat functions of the parameters θ. Informative prior distributions are defined by probabilities that are higher in certain ranges of the parameters θ where we expect the value of θ to lie, and probabilities that are lower in less likely ranges.

For the rotamer problem, we use a multinomial model, such that the likelihood takes the form

$$p(y|\theta) = \left(\frac{y_0!}{\prod_i y_i!}\right)\prod_i \theta_i^{y_i}, \tag{11}$$

where $y_0 = \sum_{i=1} y_i$, the sum of the counts of the different rotamers, and $\sum_i \theta_i = 1$. We use a *conjugate* prior distribution [a prior distribution that has a similar mathematical form to the sampling distribution (Equation 11)] called the Dirichlet distribution,

Table 6. *P-values and prediction rates for backbone-dependent rotamers*

Residue	N	1st prediction[a]	2nd prediction	3rd prediction	N bins[b]	p-value
Arg	4,572	64.7	30.0	5.4	155	0.56
Asn	5,169	70.3	23.7	6.1	222	0.50
Asp	6,575	74.9	18.9	6.2	229	0.54
Cys	1,800	70.5	25.6	3.9	113	0.66
Gln	3,833	66.2	28.6	5.2	146	0.52
Glu	6,032	61.8	32.2	6.0	162	0.46
His	2,270	71.5	24.2	4.3	149	0.65
Ile	5,792	86.7	9.6	3.7	139	0.56
Leu	8,587	73.9	25.1	1.0	166	0.65
Lys	6,176	67.1	28.2	4.7	181	0.50
Met	2,068	70.1	26.2	3.8	104	0.64
Phe	4,348	76.4	21.2	2.4	167	0.73
Ser	6,906	62.5	26.6	10.9	194	0.44
Thr	6,539	83.9	11.1	5.0	166	0.48
Trp	1,632	68.9	26.4	4.7	101	0.62
Tyr	4,042	74.2	22.6	3.2	156	0.66
Val	7,576	84.3	9.6	6.1	148	0.41
All	83,917	73.0	22.0	5.0		

[a] Prediction rates determined by the $p(r_1|\phi_a,\psi_b)$ for all side chains in the database. First prediction is made based on the highest probability r_1 rotamer from the appropriate ϕ,ψ bin; second prediction is the next highest probability; third prediction is based on the lowest probability r_1 rotamer.

[b] Number of bins with $y^{non\text{-}per} > 10$ used to calculate the p-values in the last column.

$$p(\theta) = \text{Dirichlet}(\{x_i + 1\}) = \left(\frac{\Gamma(x_0 + m)}{\prod_i \Gamma(x_i + 1)}\right)\prod_{i=1}^{m} \theta_i^{x_i}, \tag{12}$$

where $x_0 = \Sigma_{i=1}x_i$. Note that for an integer $\Gamma(n + 1) = n!$. The values of the *hyperparameters* x_i that define the prior distribution can be thought of as estimated counts for the rotamers in some sample of side chains of size x_0.[1] These estimates can come from any source, including previous data or some pooling of the present data. The total number of prior counts, x_0, can be scaled to any value to alter the dependence of the posterior distribution on the prior distribution. The larger x_0 is, the more precise the prior distribution is, and the closer the posterior density is to values near $\theta_i = x_i/x_0$, The posterior distribution that results from Equations 11 and 12 is also Dirichlet with parameters $x_i + y_i + 1$, i.e.,

$$p(\theta|y) = \left(\frac{\Gamma(x_0 + y_0 + m)}{\prod_i \Gamma(x_i + y_i + 1)}\right)\prod_i \theta_i^{x_i+y_i}. \tag{13}$$

The use of the conjugate prior distribution results in the analytical form of the posterior distribution (Equation 13) and also therefore simple expressions for the expectation values for the θ_i, their variances, covariances, and modes:

$$E(\theta_i) = \frac{x_i + y_i + 1}{x_0 + y_0 + m}$$

$$\text{mode}(\theta_i) = \frac{x_i + y_i}{x_0 + y_0}$$

$$\text{var}(\theta_i) = \frac{E(\theta_i)(1 - E(\theta_i))}{x_0 + y_0 + m + 1}$$

$$\text{cov}(\theta_i,\theta_j) = -\frac{E(\theta_i)E(\theta_j)}{(x_0 + y_0 + m + 1)}, \tag{14}$$

where m is the number of rotamers under consideration ($i = 1,2,\ldots,m$).

Bayesian simulation of the posterior distribution

Bayesian inference is frequently performed by means of simulation, especially when prior distributions and sampling distributions are complicated and do not produce an analytical posterior distribution. We use such a technique as a means for finding the best prior distribution.

To check our models against the experimental data, it is useful to make predictions of the data from the posterior distribution of the parameters. We refer to these as $y^{rep,l}$ for $l = 1,2,3,\ldots,L$ for L draws from the posterior distribution. To perform simulations, we first draw values for $\theta^{rep,l}$ from a Dirichlet distribution with parameters $x_i + y_i + 1$ and then draw $y^{rep,l}$ from the multinomial likelihood function, Equation 11. With the y^{rep} in hand, we can evaluate our models with the calculation of p-values within the Bayesian statistics framework. Because Bayesian inference is performed over the entire posterior distribution of the parameters, rather than just a point value as in classical statistics, we calculate the proportion of simulation draws where the replicated data are more extreme than the experimental data,

$$p - \text{value} = \frac{1}{L}\sum_{l=1}^{L} I\{T(y^{rep,l},\theta^{rep,l}) \geq T(y,\theta^{rep,l})\}, \tag{15}$$

where the indicator function $I = 1$ if the expression inside is true and 0 otherwise. T is the *test statistic*, which is a function of both y and θ. We use a χ^2 *discrepancy*, which is similar to the classical χ^2 goodness-of-fit measure,

$$T(y,\theta^{rep,l}) = \sum_{i=1}^{m} \frac{(y_i - E(y_i|\theta^{rep,l}))^2}{\text{var}(y_i|\sigma^{rep,l})}, \tag{16}$$

where the sum is over the different rotamer types and y is either from the experimental data or the replicated data.

Rotamer libraries derived from Bayesian statistical analysis

Backbone-independent rotamer library

For the backbone-independent rotamer library, we use informative prior distributions of the form,

$$p(\theta_{jkl|i}) = \text{Dirichlet}\left\{Ky_0\frac{\beta_{i|k}\beta_{k|j}\beta_{j|i}}{\sum_{i',j',k',l'}\beta_{i'|k'}\beta_{k'|j'}\beta_{j'|i'}} + 1\right\}. \tag{17}$$

[1] We use the notation from the following table for the counts and proportions for the prior, data, and posterior distributions:

	Prior	Data	Posterior
Counts	x_i	y_i	t_i
Proportions	a_i	b_i	q_i

K is a scale factor that affects the influence of the prior distribution on the posterior distribution by setting the value of x_0 in Equation 12 proportional to the number of side chains y_0 in the experimental data with $r_1 = i$. For larger values of K, the prior distribution exerts a stronger pull on the experimental data.

Because some of the rotamer types are not seen at all in the limited data set, we would like some estimate of their probability, beyond the crude noninformative distribution result. Higher values of K are therefore desirable. We do not know for sure how good our prior distribution is and the resulting posterior distribution (which is the goal) is dependent on it. So we test the posterior distributions using the Bayesian version of the χ^2 test by simulating draws, θ^{rep}, from the posterior distributions from a range of values for K and y from the likelihood functions. The resulting p-values indicate what fraction of the sample draws yielded a value of the test χ^2 function that were larger than the value for the actual data sample. Values near 1/2 indicate that the experimental data are exactly in the middle of the posterior distribution. Values below 0.05 and above 0.95 indicate that the data do not appear to come from the posterior distribution.

We divided the data in half and derived backbone-independent rotamer libraries from each half database as well as the full database. For values of $K \leq 0.5$, when the data and posterior distribution come from the same data set, the p-values stay within the range 0.4–0.7. Using as large a value of K as possible allows us to emphasize the prior distribution in cases where N is small. When either half database was used for the posterior distribution and the full database was used as the data, the range of p-values was lower, but they increased to more median levels as K was increased. This indicates that, for situations with less experimental data (the half databases), the posterior distribution should rely heavily on an informative prior distribution. We found a value of $K = 0.5$ to be optimal. This is a useful guideline for the more complicated case of the backbone-dependent distributions, where there is much less data per parameter than in the backbone-independent case.

Backbone-dependent rotamer library

We would like to determine a backbone-dependent library that is continuous over the full Ramachandran map and that provides good estimates for regions that are only weakly populated or even unpopulated in the PDB. We can check the model in various ways to ensure that the populated regions are represented accurately.

In our prior distribution, we postulate that

$$p(r_1 = i|\phi,\psi) \propto p(r_1 = i|\phi)p(r_1 = i|\psi). \tag{18}$$

In practice, a robust estimate of prior distribution parameters $p(r_1 = i|\phi_a) \equiv \alpha_{i|a}$ and $p(r_1 = i|\psi_b) \equiv \alpha_{i|b}$ can be obtained from a loglinear model of the data probabilities $p(r_1 = i|\phi_a,\psi_b) \equiv \beta^{per}_{i|ab}$,

$$\ln \beta^{per}_{i|ab} = \ln \alpha_{i|a} + \ln \alpha_{i|b}, \tag{19}$$

This is an overdetermined problem, because there are 36 × 36 values on the left-hand side to be expressed as a linear combination of 72 values on the right-hand side. The matrix equation can be solved by singular value decomposition (SVD), and is guaranteed to be the best solution of the overdetermined problem in the least-squares sense (Press et al., 1988).

The two-dimensional prior distribution is a Dirichlet function at each value of $\phi = \phi_a$ and $\psi = \psi_b$,

$$p(\theta_{i|ab}) = \text{Dirichlet}\left\{x_{0|ab}\frac{\alpha_{i|a}\alpha_{i|b}}{\sum\limits_{i'}\alpha_{i'|a}\alpha_{i'|b}} + 1\right\}, \tag{20}$$

where the ϕ- and ψ-dependent probabilities are determined from the SVD equations (Equations 18 and 19) with the ϕ,ψ-dependent data obtained with weighting function W^{per}. We use a value for $x_{0|ab}$ in a limited range, such that

$$x_{0|ab} = \begin{cases} 20 & Ky_{0|ab} \leq 20 \\ Ky_{0|ab} & 20 < Ky_{0|ab} < 100, \\ 100 & 100 \leq Ky_{0|ab} \end{cases} \tag{21}$$

where K is whatever proportionality constant we choose.

The posterior distribution then is also Dirichlet, as in the backbone-independent case,

$$p(\theta_{i|ab}|y) = \text{Dirichlet}\{x_{i|ab} + y_{i|ab} + 1\}, \tag{22}$$

with $y_{i|ab}$ obtained from the nonperiodic function, $W^{non\text{-}per}$. The use of $W^{non\text{-}per}$ here minimizes the effect of the periodic function in populated regions of the map. To test how well the posterior distributions represent the data in the occupied regions of the Ramachandran map, we calculated the p-values for the distributions for all side chains with the results listed in Table 6.

The full rotamer library

Given the strong dependence of the χ_1 rotamer probabilities on the backbone dihedrals ϕ and ψ and the lack of dependence of r_2 on ϕ and ψ (see Results), we present the backbone-dependent rotamer library as the conditional probabilities of χ_1 rotamers given the values of the backbone dihedrals ϕ and ψ, $p(r_1|\phi,\psi)$. The full rotamer library is formed by combining (and normalizing) the conditional backbone-independent probabilities for r_2, r_3, r_4 with $p(r_1|\phi,\psi)$,

$$p(r_1 = i, r_2 = j, r_3 = k, r_4 = l|\phi = \phi_a, \psi = \psi_b)$$

$$= \frac{E(\theta_{jkl|i})E(\theta_{i|ab})}{\sum\limits_{i'} E(\theta_{jkl|i'})E(\theta_{i'|ab})}. \tag{23}$$

χ Angle averages

χ angle averages were determined from a Bayesian analysis using a normal model for both the prior and posterior distributions. We are looking for average χ angles as a function of ϕ and ψ for all rotamer combinations, e.g., for Lys, values for $\bar{\chi}_1, \bar{\chi}_2, \bar{\chi}_3, \bar{\chi}_4 \sim \phi_a, \psi_b, r_1, r_2, r_3, r_4$, There is not sufficient information in the PDB to determine this many parameters, so we determine only a subset of these values, assuming, for instance, that $\bar{\chi}_3$ and $\bar{\chi}_3$ are independent of ϕ and ψ.

In Table 7, we list the parameters that were used in the prior and posterior distributions for the χ angle averages. Generally, we expect that each χ angle's distribution is affected only by its rotameric state and the dihedral or rotameric state on either side. This is reflected in the choice of prior distribution dependencies. Some non-neighboring dependencies are included in the posterior distributions, in case they are of some relevance in the data analysis.

Table 7. *Dependence of χ angles on backbone dihedrals and rotamer types*

Residues	Posterior distribution	Prior distribution
Cys, Ser, Val, Thr	$\bar{\chi}_1 \sim \{\phi,\psi\},r_1$	$\bar{\chi}_1 \sim \phi,r_1 + \psi,r_1$
Pro	$\bar{\chi}_1 \sim \{\phi,\psi\},r_1$ $\bar{\chi}_2 \sim r_1$	$\bar{\chi}_1 \sim \phi,r_1 + \psi,r_1$ $\bar{\chi}_2 \sim r_1$
Phe, Tyr, His, Ile, Leu, Asp, Asn, Trp	$\bar{\chi}_1 \sim \{\phi,\psi\},r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2$	$\bar{\chi}_1 \sim \phi,r_1 + \psi,r_1 + r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2$
Met, Glu, Gln	$\bar{\chi}_1 \sim \{\phi,\psi\},r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2,r_3$ $\bar{\chi}_3 \sim r_1,r_2,r_3$	$\bar{\chi}_1 \sim \phi,r_1 + \psi,r_1 + r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2 + r_2,r_3$ $\bar{\chi}_3 \sim r_2,r_3$
Arg, Lys	$\bar{\chi}_1 \sim \{\phi,\psi\},r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2,r_3,r_4$ $\bar{\chi}_3 \sim r_1,r_2,r_3,r_4$ $\bar{\chi}_4 \sim r_1,r_2,r_3,r_4$	$\bar{\chi}_1 \sim \phi,r_1 + \psi,r_1 + r_1,r_2$ $\bar{\chi}_2 \sim r_1,r_2 + r_2,r_3$ $\bar{\chi}_3 \sim r_2,r_3 + r_3,r_4$ $\bar{\chi}_4 \sim r_3,r_4$

For each term in the prior distributions, we use a normal distribution by calculating a mean and variance of the χ angle from side chains in the database with the given values of ϕ, ψ, r_1, etc. If we have two different prior distributions for a parameter, $p_1(\theta)$ and $p_2(\theta)$, we can estimate the joint prior distribution as their product, $p(\theta) \sim p_1(\theta)p_2(\theta)$. So for instance, in combining the ϕ and ψ dependencies for Cys, Ser, Val, and Thr,

$$p(\theta) \propto \exp\left(-\frac{(\theta-\mu_\phi)^2}{2\sigma_\phi^2} - \frac{(\theta-\mu_\psi)^2}{2\sigma_\psi^2}\right), \qquad (24)$$

where the values of μ and σ come from the data distribution. By completing the square, we find that

$$p(\theta) \propto \exp\left(-\frac{(\theta-\mu)^2}{2\sigma^2}\right)$$

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_\phi^2} + \frac{1}{\sigma_\psi^2}$$

$$\mu = \sigma^2\left(\frac{\mu_\phi}{\sigma_\phi^2} + \frac{\mu_\psi}{\sigma_\psi^2}\right). \qquad (25)$$

We calculate prior distributions based on the expressions in Table 7 and Equation 25 to yield a normal distribution of mean $\bar{\chi}_{prior}$ and variance σ_{prior}^2. The posterior distribution is then calculated by adding in the data with all the correct rotameric states as defined in Table 7, yielding values for $\bar{\chi}_{post}$ and σ_{post}^2. The posterior distribution is normal,

$$p(\theta|y) = N(\theta|\bar{\chi}_{post}, \sigma_{post}^2), \qquad (26)$$

where

$$\bar{\chi}_{post} = \frac{\dfrac{\bar{\chi}_{prior}}{\sigma_{prior}^2} + \dfrac{n_{data}\bar{\chi}_{data}}{\sigma_{data}^2}}{\dfrac{1}{\sigma_{prior}^2} + \dfrac{n_{data}}{\sigma_{data}^2}}$$

$$\frac{1}{\sigma_{post}^2} = \frac{1}{\sigma_{prior}^2} + \frac{n_{data}}{\sigma_{data}^2}. \qquad (27)$$

The assumption that dihedral averages depend only on the neighboring rotameric states can be tested by calculating test statistics as described above. It should be noted that σ_{data}^2 is the variance in the χ angle experimental data, whereas σ_{post}^2 is the variance in the *mean* of the χ angle, i.e., an expression of our uncertainty in the mean value, not the variance in the data from an infinitely large PDB.

Molecular mechanics calculations

We have used the program CHARMM (Brooks et al., 1983) to calculate the energies of side-chain rotamers in both the backbone-independent and backbone-dependent contexts. For all of these calculations, we have used the CHARMM22 potential energy function, which includes all atoms (polar and nonpolar hydrogens) and which has been optimized to represent a variety of intramolecular and intermolecular interactions in proteins. The backbone-independent energies were calculated by considering the atoms in a single residue fragment consisting of only N, Cα, Hα, C, and the residue side-chain.

Backbone-dependent rotamer preferences were calculated with unconstrained χ_1 (and χ_2) dihedrals starting from near the likely minima (60°, 180°, −60°) by fixing the ϕ and ψ dihedrals of the N-acetyl N′-methylamide of each amino acid with force constants of 1,000 kcal/mol and minimizing with 1,000 steps of the conjugate gradient minimizer. We placed an oxygen atom bonded to NH of each dipeptide in the position of a likely hydrogen bond acceptor. The oxygen was fixed at a distance of 3.0 Å from the backbone N and collinear with the NH bond. The dielectric constant ϵ was set to 1.0. In most cases, the minimized χ_1 values were less than 35° from the starting conformations. In others, no local minimum was found and the final χ_1 value was over 100° from the original value. If the minimized dihedral angle value was more than 60° away from the initial value, a force constant of 100 kcal/mol was applied to the χ dihedrals at their initial values, and the minimization was repeated. This ensures that the rotamer energies (after excluding the constraint energies) and calculated probabilities correspond to the correct rotamers.

To remove the influence of the backbone–backbone steric interactions and to plot the database energies and CHARMM energies, $E_i(\phi_a\psi_b)$, in a similar way, we invert the Boltzmann calculation,

$$p_{i|ab} = \frac{\exp\left(\dfrac{-E_i(\phi_a,\psi_b)}{kT}\right)}{\sum\limits_{i'}\exp\left(\dfrac{-E_{i'}(\phi_a,\psi_b)}{kT}\right)}$$

$$F_{i|ab} \equiv -kT\ln p_{i|ab}$$

$$= E_i(\phi_a,\psi_b) + kT\ln\left(\sum_{i'}\exp\left(\frac{-E_{i'}(\phi_a,\psi_b)}{kT}\right)\right). \qquad (28)$$

The value of kT was set to 1.0.

Graphical analysis was performed with the Trellis graphics module (Becker & Cleveland, 1996) of the program S-PLUS (MathSoft, 1996).

Acknowledgments

This work is supported by grants from the National Institutes of Health. R.L.D. is an NIH Post-doctoral Fellow.

References

Becker RA, Cleveland WS. 1996. *S-PLUS Trellis Graphics User's Manual*. Seattle: MathSoft, Inc. Murray Hill: Bell Labs.

Benedetti E, Morelli G, Nemethy G, Scheraga HA. 1983. Statistical and energetic analysis of side-chain conformations in oligopeptides. *Int J Peptide Protein Res 22*:1–15.

Bhat TN, Sasisekharan V, Vijayan M. 1979. An analysis of side-chain conformation in proteins. *Int J Peptide Protein Res 13*:170–184.

Bower M, Cohen FE, Dunbrack RL Jr. 1997. Homology modeling with a backbone-dependent rotamer library. *J Mol Biol 267*:1268–1282.

Bromberg S, Dill KA. 1994. Side-chain entropy and packing in proteins. *Protein Sci 3*:997–1009.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J Comp Chem 4*:187–217.

Chandrasekaran R, Ramachandran GN. 1970. Studies on the conformation of amino acids. XI. Analysis of the observed side group conformations in proteins. *Int J Pept Prot Res 2*:223–233.

Chinea G, Padron G, Hooft RW, Sander C, Vriend G. 1995. The use of position-specific rotamers in model building by homology. *Proteins Struct Funct Genet 23*:415–421.

Cody V, Duax WL, Hauptman H. 1973. Conformational analysis of aromatic amino acids by X-ray crystallography. *Int J Pept Prot Res 5*:297–308.

Compton DAC, Montero S, Murphy WF. 1980. Low-frequency Raman spectrum and asymmetric potential function for internal rotation of gaseous *n*-butane. *J Phys Chem 84*:3587–3591.

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995. Principles of protein folding—A perspective from simple exact models. *Protein Sci 4*:561–602.

Dunbrack RL Jr. 1997. The backbone-dependent rotamer library webpage. University of California San Francisco. URL: http://www.cmpharm.ucsf.edu/~dunbrack.

Dunbrack RL Jr, Karplus M. 1993. Backbone-dependent rotamer library for proteins: Application to side-chain prediction. *J Mol Biol 230*:543–571.

Dunbrack RL Jr, Karplus M. 1994. Conformational analysis of the backbone-dependent rotamer preferences of protein side chains. *Nature Struct Biol 1*:334–340.

Durig JR, Compton DAC. 1979. Analysis of torsional spectra of molecules with two internal C_{3v} rotors. 12. Low frequency vibrational spectra, methyl torsional potential function, and internal rotation of *n*-butane. *J Phys Chem 83*:265–268.

Gelin BR, Karplus M. 1979. Side-chain torsional potentials: Effect of dipeptide, protein, and solvent environment. *Biochemistry 18*:1256–1268.

Gelman A, Carlin JB, Stern HS, Rubin DB. 1995. *Bayesian data analysis*. London: Chapman & Hall.

Heringa J, Sommerfeldt H, Higgins D, Argos P. 1992. OBSTRUCT: A program to obtain largest cliques from a protein sequence set according to structural resolution and sequence similarity. *CABIOS 8*:599–600.

James MNG, Sielecki AR. 1983. Structure and refinement of penicillopepsin at 1.8 Å resolution. *J Mol Biol 163*:299–361.

Janin J, Wodak S, Levitt M, Maigret B. 1978. Conformations of amino acid side chains in proteins. *J Mol Biol 125*:357–386.

Karplus M, Parr RG. 1963. An approach to the internal rotation problem. *J Chem Phys 38*:1547–1552.

Kemp JD, Pitzer KS. 1937. The entropy of ethane and the third law of thermodynamics. Hindered rotation of methyl groups. *J Am Chem Soc 59*:276–279.

Kuszewski J, Gronenborn AM, Clore GM. 1996. Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci 5*:1067–1080.

Levitt M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J Mol Biol 226*:507–533.

Lewis PN, Momany FA, Scheraga HA. 1973. Energy parameters in polypeptides. VI. Conformational energy analysis of the N-acetyl N′-methyl amides of the twenty naturally occurring amino acids. *Israel J Chem 11*:121–152.

Marcus E, Keller DA, Shibata M, Ornstein RL, Rein R. 1996. Comparing theoretical and experimental backbone-dependent side-chain conformational preferences for linear, branched, aromatic, and polar residues. *Chem Phys 204*:157–171.

MathSoft. 1996. *S-PLUS 3.4 for Unix*. Seattle: Data Analysis Products Division, MathSoft.

McGregor MJ, Islam SA, Sternberg MJE. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol 198*:295–310.

Moult J, James MNG. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins Strict Funct Genet 1*:146–163.

Nayeem A, Scheraga HA. 1994. A statistical analysis of side-chain conformations in proteins—Comparison with ECEPP predictions. *J Protein Chem 13*:283–296.

Pitzer KS. 1940a. Chemical equilibria, free energies, and heat contents for gaseous hydrocarbons. *Chem Rev 27*:39–57.

Pitzer KS. 1940b. The vibration frequencies and thermodynamic functions of long chain hydrocarbons. *J Chem Phys 8*:711–720.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–792.

Ponnuswamy PK, Sasisekharan V. 1971. Studies on the conformation of amino acids. IX. Conformations of butyl, seryl, threonyl, cysteinyl, and valyl residues in a dipeptide unit. *Biopolymers 10*:565–582.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1988. *Numerical Recipes in C*. Cambridge, England: Cambridge University Press.

Pullman B, Pullman A. 1974. Molecular orbital calculations on the conformation of amino acid residues of proteins. *Adv Protein Chem 28*:347–526.

Šali A, Blundell TL. 1993. Comparative protein modeling by satisfaction of spatial restraints. *J Mol Biol 234*:779–815.

Sasisekharan V, Ponnuswamy PK. 1971. Studies on the conformation of amino acids. X. Conformations of norvalyl, leucyl, aromatic side groups in a dipeptide unit. *Biopolymers 10*:583–592.

Schrauber H, Eisenhaber F, Argos P. 1993. Rotamers: To be or not to be? An analysis of amino acid side-chain conformations in globular proteins. *J Mol Biol 230*:592–612.

Shakhnovich EI, Finkelstein AV. 1989. Theory of cooperative transitions in protein molecules 1. Why denaturation of globular protein is a first-order phase transition. *Biopolymers 28*:1667–1680.

Sippl MJ. 1990. Calculation of conformational ensembles from potentials of mean force: An approach to the knowledge-based predictions of local structures in globular proteins. *J Mol Biol 213*:859–883.

Sutcliffe MJ, Hayes FR, Blundell TL. 1987. Knowledge-based modeling of homologous proteins, Part II: Rules for the conformations of substituted side chains. *Protein Eng 1*:385–392.

Tuffery P, Etchebest C, Hazout S, Lavery R. 1991. A new approach to the rapid determination of protein side-chain conformations. *J Biomol Str Dynam 8*:1267–1289.

Wiberg KB, Murcko MA. 1988. Rotational barriers. 2. Energies of alkane rotamers. An examination of gauche interactions. *J Am Chem Soc 110*:8029–8038.

Zimmerman SS, Scheraga HA. 1978. Influence of local interactions on protein structure. IV. Conformational energy studies of N-acetyl-N′-methylamides of Ser-X and X-Ser dipeptides. *Biopolymers 17*:1885–1890.